統計数理 (1999)
第47巻 第1号 3-27

特集「統計基礎理論」
[原著論文]

モデル選択理論の新展開

統計数理研究所　下　平　英　寿

(受付 1998年10月6日；改訂 1998年12月15日)

要　旨

確率モデルに基づくデータ解析の有効性は，様々な分野の応用を通して示されてきた．しかし対象に関する事前知識だけから適切なモデルをひとつに決めることは一般に困難であり，データに基づいてモデルを選択する方法が必要になる．赤池は予測の観点からモデルの良さを評価するための情報量規準を提案し，データ解析におけるモデリングの重要性を説いた．現在では，情報量規準は種々のものが提案されており，使われる状況に応じてそれらを使い分ける必要がある．本稿では，推測方式に応じた情報量規準の導出を議論し，またモデル選択の一致性についても調べる．一致性はサンプルサイズが大きくなる極限での問題だが，実際のデータ解析ではサンプルサイズは有限である．従ってどの規準を使うにしてもそのサンプリングエラーを考慮して，モデル選択の信頼性（又は不確実性）を定量的に評価することが重要である．このためのいくつかの方法（ブートストラップ選択確率，モデル選択検定，多重比較によるモデルの信頼集合）を議論する．さらに，探索的モデル構築のために，予測分布の相対的な関係を直接視覚化するためのグラフィカルな方法の有効性を，重回帰の変数選択や遺伝データ解析を数値例にあげて示す．

キーワード：情報量規準，AIC，予測分布，変数選択，ベイズモデル，多重比較法．

1. はじめに

Akaike (1974) による赤池情報量規準 (Akaike Information Criterion, AIC) の導入以来，AIC最小化法をはじめとするモデル選択は，多くの応用分野で成果をあげている．特に時系列解析ではAICの実用例は数多く報告されている（赤池・北川 (1994)；北川・樋口 (1998)）．例えば，時系列 $x_1,\ldots, x_n$ を説明するモデルとして，m次の自己回帰 (AR) モデル

$$x_t = \sum_{i=1}^{m} a_i x_{t-i} + \epsilon_t\,; \qquad \epsilon_t \sim N(0,\ \sigma_m^2), \quad t = 1,\ldots,\ n,\ \text{i.i.d.} \tag{1.1}$$

を考える．次数mが未知の場合，これをデータから決めるには様々なアプローチが有り得るが，そのひとつは「平均的に良い予測を与えるm」を選ぶことである．その推定として，次のAICを最小化する $\hat{m}$ を用いればよい．

$$\mathrm{AIC}\,(m) := n \log\, \hat{\sigma}_m^2 + 2m \tag{1.2}$$

ここで，$\hat{\sigma}_m^2$ は σ_m^2 の最尤推定 (MLE) であり，またmに依存しない定数は省略してある．(1.2) と定数倍だけ異なるAICの定義もあるが，最小化には影響しない．$m = \hat{m}$ とし，モデルのパラ

メタ $\theta_m=(a_1,\ldots,a_m,\sigma_m^2)$ を MLE $\hat{\theta}_m$ で推定する事によりシステムが同定される．

この方法は情報理論的な視点から一般化されて，AR モデル以外の時系列モデルへの適用や，また時系列以外への様々な応用がなされている．しかし，次のような点に関しての疑問や反省も指摘されている．(i) MLE 以外のパラメタ推定法が用いられるときは AIC はどのように修正すればよいか？ さらに一般化して，ベイズ予測分布などの推測方式ではどうなるか？ (ii) MLE を使うとしても，AIC 以外の様々なモデル選択規準が提案されているが，どれが良いのか？ モデル選択の一致性の観点では，AIC は良くない方法なのか？ (iii) いくつかのモデルの AIC の値があまり違わない時，それらの差が「有意」といえるのか？ もし有意でなければ，選択の不確実性はどのように表現されるか？ などである．本稿では，AIC の導出を概観した後で，これらの疑問点について研究状況の解説を踏まえて議論する．より全般的なサーベイは，柴田 (1988) や竹内 ((1989), pp. 459-465) などが参考になる．

2. モデル選択法

2.1 予測分布の良さと AIC

ここでは，情報量規準の最も基本的な形である AIC の導出を振り返り，以降の節の議論にそなえる．確率変数 x の密度関数を $q(x)$ とする．本稿では簡単のため，データ $x^{(n)}=(x_t: t=1,\ldots,n)$ の各要素 x_t は互いに独立に同じ密度関数 $q(x_t)$ に従うとする．未知の $q(x)$ を表すために，$\theta\in\Theta\subset\mathcal{R}^{\dim\theta}$ をパラメタとする確率モデル $p(x|\theta)$ を考える．$x^{(n)}$ の同時密度のモデルは $p(x^{(n)}|\theta)=p(x_1|\theta)\cdots p(x_n|\theta)$ であり，この対数を取った

$$L^{(n)}(\theta):=\sum_{t=1}^{n}\log p(x_t|\theta) \tag{2.1}$$

が対数尤度である．(2.1) を最大にする $\hat{\theta}\in\Theta$ が MLE であるが，この推定の良さを次のように予測分布の良さによって考える．まだ観測していない x_{n+1} の確率密度を $p(x_{n+1}|\hat{\theta})$ によって予測するとき，これが真の密度 $q(x_{n+1})$ からどれだけ離れているかを，密度間の遠さを測る尺度である Kullback-Leibler の情報量

$$D(q\,;p(\hat{\theta})):=\int q(x_{n+1})\log q(x_{n+1})\,dx_{n+1}-\int q(x_{n+1})\log p(x_{n+1}|\hat{\theta})\,dx_{n+1} \tag{2.2}$$

によって与える (図1)．$D(q\,;p(\hat{\theta}))\geq 0$ であり，等号は二つの密度関数が一致する時に限る．(2.2) の第1項はモデル $p(x|\theta)$ に依存しない定数なので，第2項をパラメタ $\hat{\theta}$ における損失 (loss) とする．これが大きいほど，そのパラメタ値が悪いと考えられる．第2項に含まれる $\hat{\theta}=\hat{\theta}(x^{(n)})$ はデータから計算される確率変数だから，予測分布の平均的な良さ (悪さ) は損失の期待値を取って，

$$-\int q(x_1)\cdots q(x_n)\int q(x_{n+1})\log p(x_{n+1}|\hat{\theta}(x^{(n)}))\,dx_{n+1}\,dx_1\cdots dx_n \tag{2.3}$$

によって測られる．これがリスク (risk) であり，ここでは，良いモデルとは (2.3) を小さくするモデルであると考える．

AIC は (2.3) の推定量である．後で示すように $n\to\infty$ で

$$-\frac{1}{n}L^{(n)}(\hat{\theta})+\frac{1}{n}\dim\theta \tag{2.4}$$

の漸近期待値は近似的に (2.3) である．なお，x_{n+1} の予測分布の代わりに $x_{n+1},\ldots,x_{2n}$ の予測分

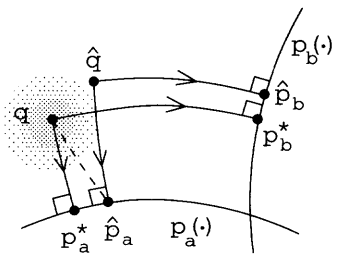

図1. 確率分布の空間におけるMLEなどの幾何的解釈 (Amari (1985)). 各点はxの確率分布を表す. モデルのパラメタを可能な範囲で動かすと点の集合ができるが, それは多様体 (曲面) によって表される. この図では互いにネストの関係にない二つのモデル $p_a(x|\theta_a)$ と $p_b(x|\theta_b)$ がある. 真の分布 $q = q(\cdot)$ からモデルを表す多様体 $p_a(\cdot) = \{p_a(\cdot|\theta_a) : \theta_a \in \Theta_a\}$ に射影した点 $p_a^* = p_a(\cdot|\theta_a^*)$ が, モデルの中でもっとも q に近い点である. 射影は, K-L 情報量 $D(q\,;\,p_a(\theta))$ の最小化として定義される. 一方, $\hat{q}$ はデータを表す経験分布で, これからモデル $p_a(\cdot)$ に射影した点 $\hat{p}_a = p(\cdot|\hat{\theta}_a)$ がMLEに相当する予測分布である. 真の分布 q からの予測分布 $\hat{p}_a$ の隔たりを $D(q\,;\,\hat{p}_a)$ で測ると, AICは (モデルによらない定数項と定数倍を除いて) その期待値の推定量になる.

布の良さを考えるときは, (2.4) を n 倍すればよい. 歴史的理由でAICは (2.4) を $2n$ 倍したものとして定義される.

竹内 (1976, 1983) に従って (2.4) の導出を概観する. モデル $p(x|\theta)$ の最適なパラメタ値 θ^* を, $D(q\,;\,p(\theta)) = \int q(x)\log q(x)\,dx - \int q(x)\log p(x|\theta)\,dx$ を最小にする $\theta \in \Theta$ と定義する. モデルが真の分布を含めば $q(\cdot) \equiv p(\cdot|\theta^*)$ であるが, 本稿ではこれは仮定しない (misspecification). また, θ^* は Θ の内点とし, モデルのなめらかさ等に関する適当な正則条件 (White (1982)) を仮定しておく.

大数の法則より, $n \to \infty$ で $-L^{(n)}(\theta)/n \to -\int q(x)\log p(x|\theta)\,dx$ が言えるので, $\hat{\theta} \to \theta^*$ である. さらに,

$$\hat{\theta} \overset{\mathrm{A}}{\sim} N(\theta^*,\, H^{-1}GH^{-1}/n) \tag{2.5}$$

のように, MLEは漸近的に正規分布に従う. ただし,

$$G := \int q(x)\frac{\partial \log p(x|\theta)}{\partial\theta}\bigg|_{\theta^*}\frac{\partial \log p(x|\theta)}{\partial\theta'}\bigg|_{\theta^*}dx\,, \qquad H := -\int q(x)\frac{\partial^2 \log p(x|\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta^*}dx$$

である. もしモデルが真の分布を含むと $G = H$ となり, これはFisher情報行列になるが, 一般には $G \neq H$ である. (2.5) は, 推定方程式

$$\sum_{t=1}^{n}\frac{\partial \log p(x_t|\theta)}{\partial\theta}\bigg|_{\hat{\theta}} = 0 \tag{2.6}$$

を θ^* の周りでテーラ展開すると中心極限定理より直ちに得られる.

さて, $L^{(n)}(\theta^*)$ を $\hat{\theta}$ の周りでテーラ展開すると, (2.6) より,

$$-L^{(n)}(\theta^*) \approx -L^{(n)}(\hat{\theta}) + \frac{n}{2}(\theta^* - \hat{\theta})'\,H(\theta^* - \hat{\theta})$$

であるから, 両辺の期待値を取って n で割ると, (2.5) に注意して,

$$-\int q(x_{n+1})\log p(x_{n+1}|\theta^*)\,dx_{n+1} \tag{2.7}$$

$$\approx -\frac{1}{n}\int q(x^{(n)})\, L^{(n)}(\hat{\theta}(x^{(n)}))\, dx^{(n)} + \frac{1}{2n}\mathrm{tr}\,(H\cdot H^{-1}GH^{-1})$$

一方，$\log p(x_{n+1}|\hat{\theta})$ を θ^* の周りでテーラ展開すると，$\int q(x_{n+1})(\partial \log p(x_{n+1}|\theta)/\partial\theta)|_{\theta^*}\, dx_{n+1} = 0$ などに注意して，(2.3) は

$$-\int q(x_{n+1}) \log p(x_{n+1}|\theta^*)\, dx_{n+1} + \frac{1}{2n}\mathrm{tr}\,(GH^{-1}) \tag{2.8}$$

と近似できる．結局 (2.7) と (2.8) を合わせて，

$$-\frac{1}{n}L^{(n)}(\hat{\theta}) + \frac{1}{n}\mathrm{tr}\,(GH^{-1}) \tag{2.9}$$

の $q(x^{(n)})$ に関する期待値は $o(n^{-1})$ の誤差で (2.3) になる．(2.9) は竹内の TIC と呼ばれ，$G \neq H$ まで考慮した，(2.4) より精密な AIC といえる．実際の計算では $\mathrm{tr}\,(GH^{-1})$ はその一致推定量などで置き換える．

正しいモデルが分かっていればモデル選択をする必要もないので，misspecification を仮定するのは妥当であり，したがって一般に $G \neq H$ である．しかし一方，あまりに不適切なモデルは自明に棄却されるので，$p(\cdot|\theta^*)$ が $q(\cdot)$ を比較的良く近似すると仮定するのも妥当である．この時 $G \approx H$ と近似する事が許されて，$\mathrm{tr}\,(GH^{-1}) \approx \dim\theta$ となり，(2.9) から (2.4) を得る．

なお，漸近論で $n \to \infty$ とするのは十分にデータがある場合を近似的に表現する数学的手段であるが，そのときモデルや真の分布の関係を上での議論のように固定して考える方法を“fixed alternatives”と呼ぶことがある．これに対して，$n \to \infty$ につれて $D(q;\, p(\theta^*))^{1/2} = O(n^{-1/2})$ のオーダでモデルを真の分布に近づけることは検定の漸近論で“local alternatives”と呼ばれる．これは，実際のデータ解析において，n が大きくなるにつれて想定するモデルのクラスが変化することのひとつの表現と言える．ここで $O(n^{-1/2})$ のオーダを用いることには必ずしも合理的な根拠があるわけではないが，このオーダで評価すると AIC の第1項と第2項の大きさのオーダが同じになるので，$n \to \infty$ としたときも両方の効果が無視されないという利点がある．Shimodaira (1997) は local alternatives におけるモデル選択を議論し，$G = H$ として良いことを示している．fixed alternatives による評価の問題点は，3節の一致性の議論でも再び触れる．

例1．（重回帰モデルの変数選択） m 個の説明変数 $z_1,\ldots,\, z_m$ の線形結合で従属変数 y を予測する．$x_t = (y_t,\, z_{1,t},\ldots,\, z_{m,t})$ とし，$z_t = (z_{1,t},\ldots,\, z_{m,t})$ を与えた時の y_t の条件付分布 $q(y_t|z_t)$ のモデル $p_m(y_t|z_t,\, \theta_m)$ を，$\theta_m = (\beta_1,\ldots,\, \beta_m,\, \sigma_m^2)$ として

$$y_t = \sum_{i=1}^{m}\beta_i z_{i,t} + \epsilon_t\,; \qquad \epsilon_t \sim N(0,\, \sigma_m^2), \quad t=1,\ldots,\, n,\ \text{i.i.d.} \tag{2.10}$$

によって与える．$q(z_t)$ は未知だが，形式的に $p_m(x_t|\theta_m) = p_m(y_t|z_t,\, \theta_m)\, q(z_t)$ と与える事により，式 (2.1) 以下の議論がそのまま成り立つ．$q(z_t)$ は θ を含まないから，推定方程式 (2.6) は

$$\sum_{t=1}^{n}\frac{\partial \log p_m(y_t|z_t,\, \theta_m)}{\partial\theta_m}\bigg|_{\hat{\theta}_m} = 0 \tag{2.11}$$

となる（坂元 他 (1983)，p. 61）．一般に $G \neq H$ であるが，$q(x_t)$ が多変量正規分布と仮定すると $G = H$ となり，TIC は AIC と同じになる (Shimodaira (1993))．また $L^{(n)}(\hat{\theta}) = -(n/2)(1 + \log 2\pi\hat{\sigma}_m^2)$ であるので，情報量規準は (1.2) を用いればよい．

m個の説明変数のいくつか z_i, $i \in \alpha \subset \{1,\ldots, m\}$ だけを使った部分モデルの方が，すべてを使ったフルモデルよりかえって (2.3) が小さく予測が良い場合がある．そこで，変数選択問題では AIC 最小化によって良い α を探す．なお，以下の例では説明変数 $z_1, z_2,\ldots$ を a, b,... のようにアルファベットで表し，予測に使う説明変数を〈·〉で囲んで表示する．例えば，〈ab〉がフルモデルとすると，〈a〉と〈b〉は互いにネストでない部分モデルである (図1)．

例 2.（AIC による AR モデルの次数選択） AR モデルは形式的に重回帰の一例と見なせる．すなわち，$z_t = (y_{t-1}, y_{t-2},\ldots, y_{t-m})$ とおくと，初期分布の影響を無視して MLE の推定方程式は漸近的に (2.11) となり，AIC は (1.2) となる．また，Akaike (1969) の FPE $= \hat{\sigma}_m^2(n+m)/(n-m)$ は，x_{n+1} の平均2乗予測誤差の推定量であるが，$n \to \infty$ で $(n/2)\log$ FPE は (1.2) で近似できる (坂元 他 (1983)，p. 148)．次数 $m_1 > m_2$ である二つの AR モデルは，図6左のように次数の大きいモデルが小さいモデルを含むネストの状況になる．

例 3.（交差確認法） $x_1,\ldots, x_n$ から x_t を除いたデータから計算した MLE を $\hat{\theta}_{-t}$ と表す．(2.3) の推定として，Cross-validation

$$-\frac{1}{n}\sum_{t=1}^{n}\log p(x_t|\hat{\theta}_{-t}) \tag{2.12}$$

が用いられる事がある．漸近的に (2.12) は (2.9) に等価である事が知られている (Stone (1977); Shibata (1989))．

2.2 色々な情報量規準

前節では MLE $\hat{\theta}$ に関する予測分布 $p(x_{n+1}|\hat{\theta}(x^{(n)}))$ の良さを調べた．ここでは MLE 等の推定量 $\tilde{\theta}$ に関する予測分布 $p(x_{n+1}|\tilde{\theta}(x^{(n)}))$ の場合 (2.2.1 節) や，ベイズの予測分布 (2.2.2 節) を例としてとりあげ，推測方式 (＝予測分布の作り方) に応じた情報量規準の導出を議論する．

データ $x^{(n)}$ から計算した x_{n+1} の予測分布を一般に $p(x_{n+1}|x^{(n)})$ と書く．これが $q(x_{n+1})$ から平均的にどれだけ離れているかを，(2.3) と全く同様に

$$-\int q(x_1)\cdots q(x_n)\int q(x_{n+1})\log p(x_{n+1}|x^{(n)})\,dx_{n+1}\,dx_1\cdots dx_n \tag{2.13}$$

をリスクとして測る．本節で扱う情報量規準は，このリスクの推定量として定式化される．予測分布の平均的な良さの推定量である情報量規準は，単に「モデル」を選択するだけでなく，(モデル，推測方式) というペアを選択するためのものである．

推測方式の構成において，さらにデータを得る方式も異なる場合には，多少の工夫が必要となる．x の一部しか観測できない不完全データの場合 (2.3.3 節) や，実験計画などによって説明変数の分布がゆがめられている場合 (2.2.4 節) についても触れることにする．

2.2.1 パラメタ推定による予測分布の場合

MLE 等のパラメタ推定 $\tilde{\theta}$ を密度関数に代入して予測分布 $p(x_{n+1}|\tilde{\theta})$ を作る場合をまず考える．ここでは，確率分布からパラメタへの汎関数 $T(\cdot)$ により，推定量 $\tilde{\theta} = \tilde{\theta}(x^{(n)})$ が $\tilde{\theta} = T(\tilde{q}(\cdot|x^{(n)}))$ と書けるとする．ただし $\delta_x(\cdot)$ を point mass として，

$$\hat{q}(x|x^{(n)}) = n^{-1}\sum_{t=1}^{n}\delta_{x_t}(x) \tag{2.14}$$

はデータ $x^{(n)}$ を表す経験分布である．この場合のリスク (2.13) の推定量の一般的な表現を

Konishi and Kitagawa (1996) は導出し，情報量規準 GIC と呼んだ．$\theta^* = T(q)$ とし，$T(\cdot)$ の q における影響関数を $T^{(1)}(x|q)$ と書くと，GIC は

$$-\frac{1}{n}L^{(n)}(\tilde{\theta})+\frac{1}{n}\int q(x)\frac{\partial \log p(x|\theta)}{\partial \theta'}\Big|_{\theta^*} T^{(1)}(x|q)\,dx \tag{2.15}$$

と書ける．ただし TIC と同様に実際の計算では第2項はその一致推定量などで置き換える．

以下では $\tilde{\theta}$ を $\sum_{t=1}^{n}\psi(x_t|\tilde{\theta})=0$ を推定関数とする M-推定量とする．例えば最尤推定では $\psi(x|\theta)=\partial \log p(x|\theta)/\partial\theta$ であり，M-推定量の一種である．$T(\cdot)$ の影響関数は

$$T^{(1)}(x|q)=M(\psi|q)^{-1}\psi(x|\theta^*);\qquad M(\psi|q)=-\int q(x)\frac{\partial\psi(x|\theta)}{\partial\theta'}\Big|_{\theta^*}dx$$

となるので，(2.15) の第2項 $\times n$ は

$$\mathrm{tr}\left(\int q(x)\,\psi(x|\theta^*)\frac{\partial \log p(x|\theta)}{\partial\theta'}\Big|_{\theta^*}dx\cdot M(\psi|q)^{-1}\right) \tag{2.16}$$

で与えられる．特に，$T(p(\theta))=\theta$ となる M-推定量に関しては，モデルが真の分布を含むと仮定すると (2.16) $=\dim\theta$ となり，結果として GIC は AIC と同じ表現になる．

GIC のように解析的な結果を与える代わりに，Ishiguro et al. (1997) の EIC や Shibata (1997) では，ブートストラップを使った (2.3) の推定量を与えている．

例 4.（罰金付き最尤法） 罰金付き尤度 (penalized likelihood) を最大にする推定量 $\tilde{\theta}_\lambda$ は，$\psi(x|\theta)=\partial(\log p(x|\theta)-\lambda k(\theta))/\partial\theta$ とおくことにより M-推定量と見なせる．ただし，$k(\theta)$ は罰金関数，$\lambda\in\mathcal{R}$ は罰金の重みを決める係数である．この時の情報量規準は (2.15) と (2.16) よりただちに計算できる．特に，$k(\theta)$ の形がデータに依存しない場合（ただし n に依存しても良い）には，RIC として Shibata (1989) で与えられている．この場合の情報量規準は，単にモデルを選ぶだけでなく，最適な λ を選ぶことにも使われる．

2.2.2 ベイズ予測分布の場合

パラメタ θ の事前密度を $p(\theta)$ とする．データ $x^{(n)}$ を与えた時の θ の事後密度は $p(\theta|x^{(n)})\propto p(x^{(n)}|\theta)\,p(\theta)$ である．これから x_{n+1} の予測分布を，

$$p(x_{n+1}|x^{(n)})=\int p(x_{n+1}|\theta)\,p(\theta|x^{(n)})\,d\theta \tag{2.17}$$

で与える．Konishi and Kitagawa (1996) は，このベイズ予測分布に関する (2.13) の推定量を

$$-\frac{1}{n}\sum_{t=1}^{n}\log p(x_t|x^{(n)})+\frac{1}{n}\mathrm{tr}\,(GH^{-1}) \tag{2.18}$$

と与えた．

Shimodaira (1998b) によれば，(2.18) は MLE $\hat{\theta}=\hat{\theta}(x^{(n)})$ を使って，

$$-\frac{1}{n}L^{(n)}(\hat{\theta})+\frac{1}{2n}(\mathrm{tr}\,(GH^{-1})+\dim\theta) \tag{2.19}$$

とも近似できる．これと (2.9) を比べるとわかるように，$\Delta=(\mathrm{tr}\,(GH^{-1})-\dim\theta)$ とすると，MLE とベイズの予測分布との良さの差は，$\Delta/2n+o(n^{-1})$ である．Δ はモデルの埋め込み mixture-曲率に関係した量で，図2のように曲がったモデルの「外側」に q があれば $\Delta<0$ で，「内

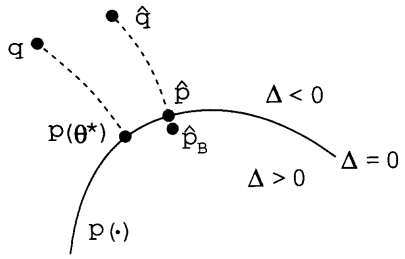

図2. モデルの曲率. MLEによる予測分布 $\hat{p}=p(\cdot|\hat{\theta})$ とベイズ予測分布 $\hat{p}_B=p(\cdot|x^{(n)})$ に関するモデルの良さ (2.3) の差は，$\Delta/2n+o(n^{-1})$ である．$\Delta>0$ ではベイズ予測分布の方が良いが，$\Delta<0$ ではMLEの方が良い．事後分布 $p(\theta|x^{(n)})$ を使って $p(\cdot|\theta)$ の平均を取ったものが $\hat{p}_B$ だから，漸近的に，$\hat{p}$ から少しモデルの「内側」にずれたところに $\hat{p}_B$ はある．このことから，この図のように q が $\Delta<0$ にあれば $\hat{p}_B$ は $\hat{p}$ よりも q から遠ざかる傾向があるのが分かる．

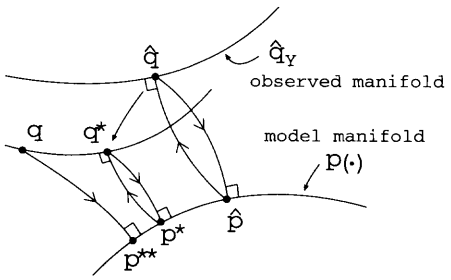

図3. 不完全データのMLEの幾何的解釈 (下平 (1992 a, 1992 b) ; Shimodaira (1994))．各点は $x=(y,z)$ の確率分布を表す． モデルを表す多様体 $p(\cdot)$ のほかに，データ $y^{(n)}$ も多様体 $\hat{q}_Y=\{q_X: \int q_X(y,z)\,dz=\hat{q}(y|y^{(n)})\}$ で表される．MLE $\hat{\theta}(y^{(n)})$ は，二つの多様体間のK-L情報量を最小にするモデル上の点として与えられる．すなわち，$D(\hat{q}\,;\,p(\hat{\theta}))=\min_{q_X\in\hat{q}_Y}\min_{\theta\in\Theta}D(q_X\,;\,p(\theta))$ であり，これは (2.20) の最大化によって得られるMLEと等価になる．二つの多様体のなす「角度」によって，$\mathrm{tr}\,(I_{Z|Y}I_Y^{-1})$ の大きさが決まる．直角に近ければ小さな値になるが，ほぼ平行ならば非常に大きな値になり得る．

側」ならば $\Delta>0$ である．もしちょうどモデルの上に q があれば $\Delta=0$ であり，Komaki (1996) によれば良さの差は $O(n^{-2})$ である．

興味深いことに，Δ はTICとAICの差であるとの解釈もできる．ただし，TICの第2項をその一致推定量で置き換える効果は無視する．Δ がモデルの曲率にmisspecificationの程度を乗じたものになるという解釈は，甘利 (私信) による．

2.2.3 不完全データの場合

完全データ $x=(y,z)$ のうち，y しか観測出来ず，z が観測できない確率変数または欠測データである場合を考える (図3)．例えば，時系列の状態空間モデルではシステムの状態は z になる．また，ベイズモデルでは z はパラメタ，θ はハイパーパラメタと見なされる．

完全データのモデル $p(x|\theta)$ を与えると，$p(y|\theta)=\int p(y,z|\theta)\,dz$ に関する対数尤度

$$L_Y^{(n)}(\theta)=\sum_{t=1}^{n}\log p(y_t|\theta) \tag{2.20}$$

を最大にする MLE $\hat{\theta} = \hat{\theta}(y^{(n)})$ は，EM アルゴリズムなどによって計算される．(2.3) の x を y に置き換えることによってこれまでの議論がすべて適用できて，例えば y_{n+1} の予測分布 $p(y_{n+1}|\hat{\theta}(y^{(n)}))$ の平均的な良さを測るための情報量規準は

$$-\frac{1}{n}L_Y^{(n)}(\hat{\theta})+\frac{1}{n}\dim\theta \tag{2.21}$$

となる．これは不完全データ $y^{(n)}$ に関する AIC であり，ベイズモデルでは ABIC と呼ばれることもある．

ところが，x_{n+1} の予測分布 $p(x_{n+1}|\hat{\theta}(y^{(n)}))$ の平均的な良さは

$$-\int q(y_1)\cdots q(y_n)\int q(x_{n+1})\log p(x_{n+1}|\hat{\theta}(y^{(n)}))\,dx_{n+1}\,dy_1\cdots dy_n \tag{2.22}$$

であり，これの推定量は AIC を修正して，

$$-\frac{1}{n}L_Y(\hat{\theta}(y^{(n)}))+\frac{1}{n}\mathrm{tr}\,(I_X I_Y^{-1}) \tag{2.23}$$

とする必要があることを Shimodaira (1994) は示した．ただし，

$$I_X(\theta)=-\int p(x|\theta)\frac{\partial^2\log p(x|\theta)}{\partial\theta\partial\theta'}dx,\qquad I_Y(\theta)=-\int p(y|\theta)\frac{\partial^2\log p(y|\theta)}{\partial\theta\partial\theta'}dy$$

はそれぞれ x と y に関する θ の Fisher 情報行列であり，(2.23) では $I_X = I_X(\hat{\theta})$, $I_Y = I_Y(\hat{\theta})$ などとする．もし $x \equiv y$ なら $I_X = I_Y$ となって，(2.23) は AIC になる．$x \not\equiv y$ の場合は $I_{Z|Y} = I_X - I_Y$ とおくと，$\mathrm{tr}\,(I_X I_Y^{-1}) = \dim\theta + \mathrm{tr}\,(I_{Z|Y} I_Y^{-1})$ と書けるので，(2.23) は AIC に相当する (2.21) に $n^{-1}\mathrm{tr}\,(I_{Z|Y} I_Y^{-1})$ を加えたものになる．これは，観測できない変数 z に関する対数尤度のバイアス補正項である．これが大きな値をとる場合は，一見すると尤度が良くても，z に関する不確定性が大きい．

2.2.4 説明変数の分布が変化する場合

回帰モデル $p(y|z,\theta)$ に関して，説明変数 z の分布がデータを観測した時と，予測分布の良さを評価する時で異なる場合，やはり AIC を修正する必要がある．例えば，説明変数の観測データにおける分布を $q_0(z)$，母集団における分布を $q_1(z)$ とし，実験計画などの理由で q_1 が q_0 とは異なるとする．ただし $q(y|z)$ は変化しない．このような場合，MLE 等の推定量 $\tilde{\theta}(x^{(n)})$ による予測分布の平均的な良さは，

$$-\int q_0(x_1)\cdots q_0(x_n)\int q_1(x_{n+1})\log p(y_{n+1}|z_{n+1},\,\tilde{\theta}(x^{(n)}))\,dx_{n+1}\,dx_1\cdots dx_n \tag{2.24}$$

で測られる．ただし，$q_0(x) = q(y|z)\,q_0(z)$，$q_1(x) = q(y|z)\,q_1(z)$ などとした．

Shimodaira (1998b) では $\tilde{\theta}$ として以下に述べる重み付き MLE を考え，そのときの (2.24) の推定量を与えた．重み関数 $w(z)$ で対数尤度を重み付けした

$$L_w(\theta) := \sum_{t=1}^{n} w(z_t)\log p(y_t|z_t,\,\theta) \tag{2.25}$$

を最大にするパラメタ値として重み付き MLE $\hat{\theta}_w \in \Theta$ を定義する (図4)．ここで重み $w_1(z) := q_1(z)/q_0(z)$ を用いた "importance sampling" を適用すると，(2.24) の $q_1(x_{n+1})$ に関する積分は $q_0(x_{n+1})\,w_1(z_{n+1})$ に関する積分に置き換えられる．あとは GIC と同様の漸近展開を行うと (2.24) の推定量として，

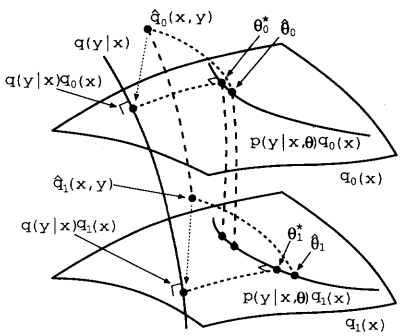

図 4. 重み付き MLE の幾何的解釈 (Shimodaira (1998 b)). データを表す経験分布 $\hat{q}(y, z)$ からモデル多様体 $\{p(y|z, \theta)\, q_0(z) : \theta \in \Theta\}$ に射影した点が MLE $\hat{\theta}$ の予測分布. 経験分布を重み関数 $w(z)$ でシフトした $\hat{q}_w(y, z) \propto w(z)\, \hat{q}(y, z)$ からモデルに射影したのが, 重み付き MLE $\hat{\theta}_w$ の予測分布. 特に $w_1(z)$ を重み関数に選んだものを, $\hat{q}_1(y, z)$, $\hat{\theta}_1$ 等と書く. この図で, $q_0(z)$ は多様体 $\{q_X(y, z) : \int q_X(y, z)\, dy = q_0(z)\}$ を表し, $q(y|z)$ は多様体 $\{q_X(y, z) : q_X(y, z)/\int q_X(y, z)\, dy = q(y|z)\}$ を表す. $q_0(z)$ と $q(y|z)$ は互いに直交している. 計量が各 foliation $q_i(z)$ で異なることより, 一般に $\hat{\theta}_0 \neq \hat{\theta}_1$ である.

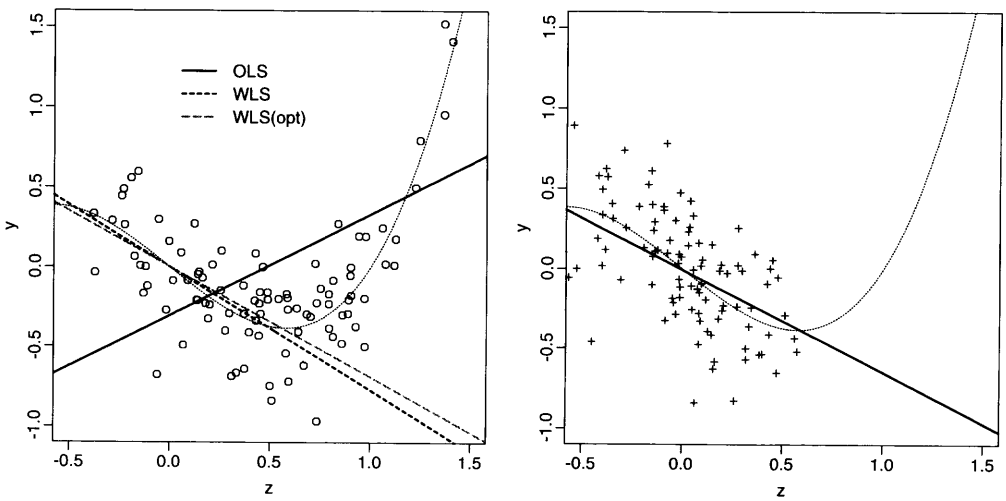

図 5. 多項式回帰の数値例. [左図] $z \sim N(0.5, 0.5^2)$; $y = -z + z^3 + \epsilon$; $\epsilon \sim N(0, 0.3^2)$ に従う $n = 100$ のデータを生成し, 1 次の多項式回帰を OLS ($\lambda = 0$) で当てはめた (実線). [右図] $z \sim N(0, 0.3^2)$ から $n = 100$ の「未来」のデータを生成し OLS を当てはめた. 左図の点線は, 過去のデータだけを使い WLS ($\lambda = 1$) で未来の回帰を推定したもので, これは右図の未来の OLS に漸近的に収束する. なお, 左図の破線は, さらに重みの最適化をした WLS ($\hat{\lambda} = 0.77$) である.

$$-\frac{1}{n} L_1(\hat{\theta}_w) + \frac{1}{n} \mathrm{tr}\,(J_w H_w^{-1}) \tag{2.26}$$

が得られる. ただし, $w_1(z)$ を用いた $L_w(\theta)$ を $L_1(\theta)$ と書き, $\theta_w^* = \mathrm{plim}_{n\to\infty} \hat{\theta}_w(x^{(n)})$ とおいて

$$J_w = \int q_0(x)\, w_1(z)\, w(z) \frac{\partial \log p(y|z,\theta)}{\partial \theta}\Big|_{\theta_w^*} \frac{\partial \log p(y|z,\theta)}{\partial \theta'}\Big|_{\theta_w^*} dx$$

$$H_w = -\int q_0(x)\, w(z) \frac{\partial^2 \log p(y|z,\theta)}{\partial \theta \partial \theta'}\Big|_{\theta_w^*} dx$$

とする．特に正規回帰では $\hat{\theta}_w$ は WLS になり $\mathrm{tr}(J_w H_w^{-1})$ の推定の計算もハット行列の対角項を使って容易に出来る．

あるモデル $p(\cdot)$ を使うとして，(2.24) を小さくする $w(z)$ を選びたい．$q \notin p(\cdot)$ で十分大きな n では，$w(z) = w_1(z)$ と選ぶのが (2.24) をもっとも小さくすることが，consistency の観点から示せる．一方，$q \in p(\cdot)$ では，$w(z) \equiv 1$ と選ぶのが，efficiency の観点から最も良い．現実のデータでは $q \notin p(\cdot)$ と考えて良いが n も有限なので，(2.26) を小さくする $w(z)$ を選ぶ．例えば，$w(z) = w_1(z)^\lambda$，$\lambda \in [0,1]$ と置いて，(2.26) を最小にする $\hat{\lambda}$ を選ぶことが考えられる (図5)．モデルの候補が複数ある場合は，$w(z)$ の選択とモデルの選択を (2.26) の最小化によって同時に行う．

2.3 ベイズ的方法

ベイズ流のモデル選択では，モデル $p(x|\theta)$ と，そのパラメタの事前分布 $p(\theta)$ が与えられると，データ $x^{(n)}$ の出現しやすさを表す

$$p(x^{(n)}) = \int p(x^{(n)}|\theta)\, p(\theta)\, d\theta \tag{2.27}$$

の値を大きくするモデルが良いとされる．これは，事前分布において各モデルが等確率としたときの事後分布に比例し，二つのモデルにおける (2.27) の比は，ベイズファクターと呼ばれる．もし事前分布において各モデルが等確率でないと考えられる場合には，それを (2.27) に掛けたものを大きくするモデルが良いとされる．なお 2.2.2 節の方法は，ベイズ予測分布の良さを非ベイズのリスクの立場で評価するもので，この節でのベイズ的方法とは異なる原理に基づいていることに注意する．

(2.27) で与えられる $p(x^{(n)})$ はパラメタを持たない「モデル」とみなせて，ベイズ的方法はその対数尤度 $\log p(x^{(n)})$ を大きくするモデルを選んでいるとも言える．もし，事前分布にハイパーパラメタ ϕ があって $p(\theta|\phi)$ と書ける場合は，(2.27) は $p(x^{(n)}|\phi)$ と書け，ϕ をパラメタと見なす尤度関数になっている．経験ベイズの立場では，ϕ を MLE $\hat{\phi}$ で置き換えるので，$p(x^{(n)}|\hat{\phi})$ の良さは，ϕ の推定に関する AIC (これを ABIC と呼ぶ)

$$-\log p(x^{(n)}|\hat{\phi}) + \dim \phi \tag{2.28}$$

によって測られる．特別な場合には (2.27) の積分が解析的に得られて，(2.28) は比較的容易に計算できる．下平 (1997) では，一種のシミュレーション技法である Markov chain Monte Carlo (MCMC) 法を利用して，(2.27) の積分を行う方法を解説している．

しかし n が十分に大きい漸近的な状況では，θ の関数として

$$p(x^{(n)}|\theta) \approx p(x^{(n)}|\hat{\theta}) \exp\left(-\frac{n}{2}(\theta - \hat{\theta})' H (\theta - \hat{\theta})\right)$$

であることを利用すると，事前分布によらずに

$$-\log p(x^{(n)}) \approx -\sum_{t=1}^{n} \log p(x_t|\hat{\theta}) + \frac{\log n}{2} \dim \theta \tag{2.29}$$

と $o(\log n)$ の誤差で近似できる．ただし，n が増大しても $\dim\theta$ が固定されていると仮定している．(2.29) はパラメタの事前分布 $p(\theta)$ に依存しない表現になっているので，$\hat{\theta}$ は MLE でも，事後分布における θ の最頻値でも良い．また，(2.28) の $\dim\phi$ も $O(1)$ なので無視して良い．(2.29) の右辺はモデル選択規準のひとつで，Schwarz (1978) の BIC と呼ばれる．これは，2段階符号化の「符号長」として得られる Rissanen (1987) の Minimum Description Length (MDL) に，$O(\log n)$ の項まで等しい (久保木 (1993)；山西 (1996))．

以下に示すように (2.29) は逐次予測の平均的な良さの推定として解釈出来る (韓 (1990))．$p(x^{(n)})=\prod_{t=1}^{n}p(x_t|x^{(t-1)})$ と分解できる事から，

$$(2.30)\quad -\int q(x^{(n)})\log p(x^{(n)})\,dx^{(n)}=-\sum_{t=1}^{n}\int q(x^{(t-1)})\int q(x_t)\log p(x_t|x^{(t-1)})\,dx_t dx^{(t-1)}$$

と書ける．データを観測する前に $p(x^{(n)})$ によって $x^{(n)}$ の予測分布を与えると考えると，その平均的な良さは (2.30) の左辺になり，(2.29) はその推定と見なせる．そして (2.30) の右辺は x_t を $x^{(t-1)}$ から逐次的にベイズ予測するときの $p(x_t|x^{(t-1)})$ の平均的な良さを $t=1,\ldots,n$ まで足し合わせたものである．また (2.8) の $1/2n$ の項を考慮すると，(2.19) よりベイズ予測分布の (2.13) と MLE の (2.3) はほぼ等しい事から，(2.30) の右辺には $\sum_{t=1}^{n}(1/2t)\approx(1/2)\log n$ の項が出て来る事も理解できる．

AIC は $x^{(n)}=(x_1,\ldots,x_n)$ を知っている場合の x_{n+1} の予測の良さを測ろうとしていたのに対し，BIC は $x^{(n)}$ を知る前の $x^{(n)}$ の予測の良さを測ろうとしている．BIC は $x^{(n)}$ を符号化して伝送する時の受け手の立場で予測を考えた規準である．この時送り手は $x^{(n)}$ を知っているから規準の計算に $x^{(n)}$ を使うことは矛盾しない．

3. モデル選択の一致性

候補となる各モデルを $p_\alpha(\cdot)$ で表し，その添字 α の集合を $\mathcal{M}$ とする．この中で一番良いモデルは (2.3) で与えたリスクを最小にするものと定義され，その添字を $\alpha^*\in\mathcal{M}$ と書く．これに対して，AIC 等の情報量規準を最小にするモデルを $\hat{\alpha}\in\mathcal{M}$ と書く．$\hat{\alpha}$ は α^* の推定と考えられるが一般には $\hat{\alpha}=\alpha^*$ とは限らない．ここでは，モデル選択の(弱)一致性 (consistency) とは，$n\to\infty$ で $\Pr(\hat{\alpha}=\alpha^*)\to 1$ が成り立つことと定義する．しばしば，「AIC は一致性を持たないが，BIC は一致性を持つので，BIC の方が良い」などと言われるが，この表現は誤解を招きやすい．結論から言うと，一致性の議論自体にあまり意味が無いし，ほとんどの場合に AIC も BIC と同様に一致性を持つことが言える．一致性が言えるための条件を本節では議論する．

各モデル $p_\alpha(x|\theta_\alpha)$ が真の分布 $q(x)$ を必ずしも含んでいない misspecification の状況を本稿では扱っている．2.1 節で与えたように，モデル $p_\alpha(x|\theta_\alpha)$ の最適なパラメタ値 θ_α^* は

$$D_\alpha(\theta_\alpha)=-\int q(x)\log p_\alpha(x|\theta_\alpha)\,dx$$

を最小にする点であるが，この最小値を $D_\alpha^*=D_\alpha(\theta_\alpha^*)$ と書く．モデルによらない定数項を除いて D_α^* は真の分布とモデルの隔たり $D(q;p_\alpha(\theta_\alpha^*))$ である．(2.8) より (2.3) は $D_\alpha^*+(1/2n)\dim\theta_\alpha$ と近似出来るので，$n\to\infty$ では D_α^* の小さいモデルが α^* となる．もし与えられた候補 $\alpha\in\mathcal{M}$ のなかに真のモデルが含まれていれば，それは D_α^* を最小にする．また，D_α^* を最小にするモデルが複数ある場合には，そのなかで $\dim\theta_\alpha$ の小さいものが $n\to\infty$ では α^* になる．

モデル選択規準としては，適当な数列 c_n を使って，

$$(3.1)\qquad \mathrm{IC}_\alpha:=-L_\alpha^{(n)}(\hat{\theta}_\alpha)+c_n\dim\theta_\alpha$$

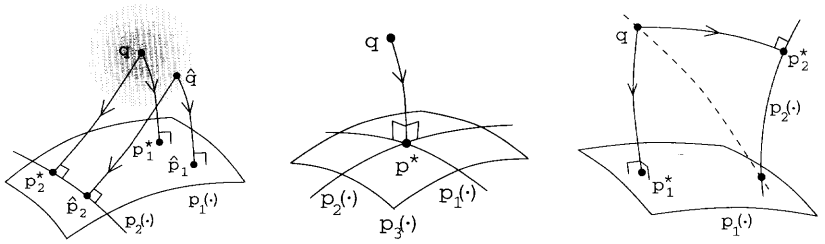

図6. 確率分布の空間におけるモデルと真の分布の位置の3パターン. [左図] 一般の位置(ネスト). なお, 図1は一般の位置でノンネスト. [中図] 最適分布が一致. [右図] 各モデルまでのK-L情報量が一致.

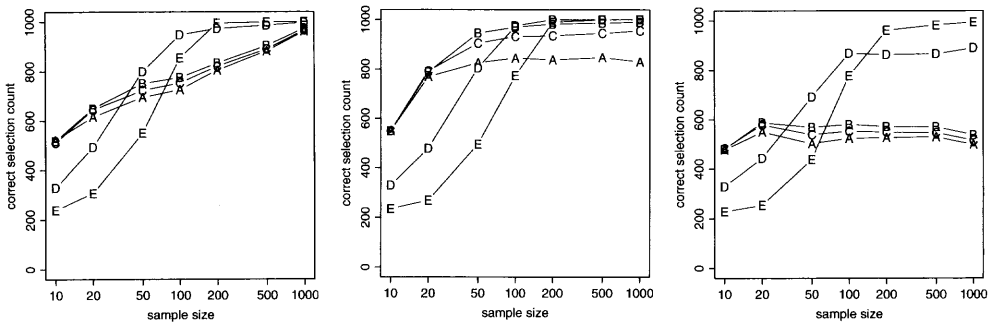

図7. シミュレーションで各規準が最も良いモデルを選んだ回数(1000回中). サンプルサイズ n を10から1000まで増やした. A: $c_n = 1$ (AIC), B: $c_n = (1/2)\log n$ (BIC), C: $c_n = n^{0.1}$, D: $c_n = n^{1/2}$, E: $c_n = n^{0.6}$. 例1で $z_{i,t}$ と ϵ_t がすべて独立に $N(0,1)$ に従うとして, データを生成した. $m = 3$ としパラメタ $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ とモデルの候補の集合 $\mathcal{M}$ を次のように3通り与えた. ただし, これらは図6の3パターンにそれぞれ対応する. [左図] $\boldsymbol{\beta} = (1.0, 0.9, 0.1)$, $\mathcal{M} = \{\langle a\rangle, \langle bc\rangle, \langle\rangle\}$. [中図] $\boldsymbol{\beta} = (1, 1, 0)$, $\mathcal{M} = \{\langle a\rangle, \langle ac\rangle, \langle\rangle\}$. [右図] $\boldsymbol{\beta} = (1, 1, 0)$, $\mathcal{M} = \{\langle a\rangle, \langle bc\rangle, \langle\rangle\}$. いずれの場合も $\langle\rangle$ は3個のモデルの中で D^* が一番大きく, 十分大きな n では選択されるのは残る2個のどちらかになる.

の形に書けるものを考える. AICなら $c_n = 1$, BICなら $c_n = (1/2)\log n$ である. 候補となる二つのモデル $p_\alpha(x|\theta_\alpha)$, $\alpha = 1, 2$ と真の分布 $q(x)$ の関係を3通りに分類して, 一致性が言えるための c_n の条件を以下では調べる.

モデルは現実を表現するひとつの近似に過ぎないと考えれば, 候補に与えたモデルはすべてmisspecificationになっていると考えるのが自然であり, このとき二つのモデルについて一般には $D_1^* \neq D_2^*$ である. このような状況のとき, 真の分布は一般の位置にあると言う(図6左, 又は図1). 漸近的に $-L_\alpha^{(n)}(\hat{\theta}_\alpha)/n \to D_\alpha^*$ が言えることに注意すると, $D_1^* \neq D_2^*$ の時(3.1)が一致性を持つ必要十分条件は $c_n = o(n)$ である. したがって, $c_n = 0$ としてもよいし, またAICもBICと同様に一致性を持つ(図7左). 以下, $c_n = o(n)$ を仮定する.

あまり現実的では無いが, 仮に $D_1^* = D_2^*$ である場合には, $\dim \theta_\alpha$ の小さい方が漸近的に良いモデルである. ひとつの状況は $p_1(x|\theta_1^*) \equiv p_2(x|\theta_2^*)$ となる場合である. (図6中では, 3つのモデルの $p_\alpha^* = p_\alpha(x|\theta_\alpha^*)$ は一致している). このような場合, 二つのモデルの $L_\alpha^{(n)}(\hat{\theta}_\alpha)$ の差が確率的に有界なので, $c_n \to \infty$ ならば一致性を持つ(Findley (1991)). したがって, AICは一致性を持たないがBICは一致性を持つ(図7中).

モデル候補の幾つかが真の分布を含むような場合はまさにこの状況であり, これがしばしば

BIC の AIC に対する優位性の根拠とされる．例えば長岡 他 (1992) ではモデルがネスト構造の場合が議論されている．Hannan (1980) は，真のモデルが存在する場合に一致性の条件 $c_n \to \infty$ を与えている．なお，$\hat{\alpha}$ が真のモデルに収束する確率が 1 である（強一致性）には $c_n = c \log \log n$, $c > 1$ とすればよい事を Hannan and Quinn (1979) が重複対数の法則より示している．このように真の分布が候補に含まれる場合の一致性の議論が文献では一般的であったが，本稿で扱っているのはすべてのモデルが間違っているかもしれないという状況での一致性である．

次のような特殊な状況では，BIC も一致性を持たない．図 6 右のように，$D_1^* = D_2^*$ であるが $p_1(x|\theta_1^*) \equiv p_2(x|\theta_2^*)$ とすると，$L_a^{(n)}(\hat{\theta}_a)$ の差が $O_p(n^{1/2})$ となるので，一致性を持つには $c_n/n^{1/2} \to \infty$ が必要十分になることが Shimodaira (1993)，下平 (1995) などで示されている．図 7 右をみると，BIC も AIC と同様に 1/2 の確率でしか 1 番良いモデルを選べない．同様の例は Findley (1991) でも示されている．このような場合でも，$c_n = n^{0.6}$ などとすれば一致性を持つ．しかしこうすると，n が小さい時には誤り確率が増える．

現実のデータでは $D_1^* \neq D_2^*$ と考えられるから，どのモデル選択規準を使っても，一番良いモデルを選ぶ確率は 1 に収束する（図 6 左）．特殊な場合には AIC は一致性を持たないことがあるが（図 6 中），別の特殊な場合には BIC も一致性を持たない（図 6 右）．実際には n は有限であるから，このようにモデルを固定して $n \to \infty$ とする一致性の議論はあまり意味が無い．Shibata (1981) では，真の分布がどのモデルにも含まれないとしたまま，n と共にモデルの次元も増やしていくような状況を考え，このときに予測の意味で AIC の最適性を示した（柴田 (1988)）．また，Shimodaira (1997) では，n と共に真の分布を各モデルに $O(n^{-1/2})$ のオーダで近づける状況を議論している．この場合には $c_n \to \infty$ となる BIC などは，考察しているモデルの中で一番小さいものを選んでしまう確率が 1 に収束してしまうので，AIC より悪いパフォーマンスを示す．一方，記述長が意味を持つ符号理論の状況では，むしろ BIC が自然な規準となる．

4. モデル選択の信頼性

有限のサンプルサイズ n では，AIC 最小モデル（もしくは (3.1) など他の規準を最小にするモデル）を選んでもそれが (2.3) の意味で最も良いモデルであるとは言えず，モデル選択には不確実性がある．したがって選択の信頼性を評価する事が重要になる．

ここでは，$\hat{\alpha}$ が α^* であることはどれくらい確からしいか？，もしくは，各モデル $\alpha \in \mathcal{M}$ が α^* である可能性はどれくらいあるのか？　という問題に定量的に答える方法を議論する．具体的には，ブートストラップ選択確率 (4.1 節)，モデル選択検定 (4.2 節)，多重比較による信頼集合 (4.3 節) について説明する．これらはいずれも，各モデル $\alpha \in \mathcal{M}$ に一種の p-値である信頼性の指標を与え，$\alpha \in \mathcal{M}$ の中で「どのモデルがもっとも良いか？」に答える．これに対して，古典的な Cox 検定や，フルモデルに対して各モデルを仮説検定する方法は，「どのモデルが正しいか？」に答えるものである．

このようにしてモデル選択の信頼性を考慮すると，$\hat{\alpha}$ をひとつだけ選ぶのではなく，比較的良いモデルを複数選ぶ方式が得られ，より良いモデルを誤って見落としてしまう可能性を減少させる．しかし，非常に多くのモデルが良いモデルとして選ばれてしまった場合などは，その解釈が重要になる．これは 5 節で議論する．

例 5.（HALD のセメントデータ） セメント発熱量 y と，説明変数として 4 種の主成分の混合率 z_i, $i = 1,\ldots,4$ のデータ ($n = 13$) を Draper and Smith ((1981), p. 629) より取った．例 1 に従いすべての説明変数の組み合わせ $2^4 = 16$ 個のモデルを $\mathcal{M}$ の要素とし，モデルの良さを評

表1. HALDデータのAICとp-値.

順位	α	AIC_α	$P_\alpha^{(L)}$	$P_\alpha^{(M)}$	$P_\alpha^{(B)}$	$P_\alpha^{(S)}$
1	<abd>	64	–	0.972	0.180	0.863
2	<abc>	+0.04	0.493	0.938	0.256	0.796
3	<ab>	+0.45	0.438	0.911	0.255	0.290
4	<acd>	+0.75	0.382	0.924	0.158	0.376
5	<abcd>	+1.97	0.088	0.451	0.002	–
6	<ad>	+3.77	0.169	0.540	0.094	0.055
7	<bcd>	+5.60	0.136	0.353	0.052	0.018
8	<cd>	+14.88	0.019	0.074	0.004	0.000
9	<bc>	+26.06	0.001	0.000	0.000	0.000
10	<d>	+33.88	0.000	0.000	0.000	0.000
11	<b>	+34.20	0.000	0.000	0.000	0.000
12	<bd>	+35.66	0.000	0.000	0.000	0.000
13	<a>	+38.55	0.000	0.000	0.000	0.000
14	<ac>	+40.14	0.000	0.000	0.000	0.000
15	<c>	+44.09	0.000	0.000	0.000	0.000
16	<>	+46.47	0.000	0.000	0.000	0.000

注) モデル $\alpha \in \mathcal{M}$ はAICの小さい順に並べてある. AIC_α はモデル α のAICで, "+"はAIC最小モデル $\hat{\alpha}$ からの差を表す. $P_\alpha^{(L)}$ は各 $\alpha \in \mathcal{M}$ の $\hat{\alpha}$ に対するLinhartのモデル選択検定のp-値(4.2節), $P_\alpha^{(M)}$ は多重比較によるp-値(4.3節), $P_\alpha^{(B)}$ はブートストラップ選択確率(4.1節), $P_\alpha^{(S)}$ はフルモデルに対する尤度比検定のp-値.

表2. BOSTONデータにおける上位20個のモデルのAICとp-値.

順位	α	AIC_α	$P_\alpha^{(L)}$	$P_\alpha^{(M)}$	$P_\alpha^{(B)}$	$P_\alpha^{(S)}$
1	<afm>	−156	–	0.994	0.487	0.000
2	<akm>	+1.72	0.461	0.986	0.381	0.000
3	<ahm>	+18.67	0.180	0.774	0.069	0.000
4	<agm>	+24.76	0.098	0.559	0.006	0.000
5	<adm>	+25.32	0.110	0.570	0.016	0.000
6	<alm>	+32.26	0.055	0.256	0.002	0.000
7	<ajm>	+35.31	0.028	0.061	0.000	0.000
8	<abm>	+38.42	0.020	0.122	0.000	0.000
9	<aim>	+39.51	0.016	0.104	0.000	0.000
10	<acm>	+40.72	0.013	0.040	0.000	0.000
11	<aem>	+40.94	0.013	0.048	0.000	0.000
12	<klm>	+46.86	0.079	0.441	0.022	0.000
13	<fjm>	+48.73	0.019	0.280	0.005	0.000
14	<jkm>	+53.99	0.028	0.140	0.000	0.000
15	<fkm>	+54.11	0.023	0.331	0.004	0.000
16	<flm>	+59.78	0.027	0.299	0.006	0.000
17	<hjm>	+60.70	0.023	0.227	0.002	0.000
18	<dkm>	+62.01	0.023	0.177	0.001	0.000
19	<bkm>	+62.69	0.018	0.148	0.000	0.000
20	<ekm>	+64.75	0.013	0.078	0.000	0.000

注) $P^{(S)}_{\langle \mathrm{afm} \rangle} = \Pr\{\chi^2_{10} > 133.8\} = 0.000$ であり, 残りのモデルはこれより小さい.

価した結果を表1に示す. $\hat{\alpha} = \langle \mathrm{abd} \rangle$ であるが, 他の $\alpha \in \mathcal{M}$ も比較的高いp-値を示すものが多く, どれが α^* であるかは必ずしも明白でない. このことは, 上位のモデルではAICの差が小さいことからも読み取れる.

例6. (BOSTONの住宅価格データ) 住宅価格とその共変量のデータをBelsley et al. ((1980), p. 244) より取った. $n = 506$ の各地域において, 住宅価格の中央値の対数を取ったも

のを y とし，$m=13$ 個の説明変数は，各地域における犯罪率 (z_1=a)，平均部屋数 (z_6 = f)，教師数：生徒数の比 (z_{11} = k)，社会階層の構成比 (z_{13} = m) などである．2^{13} 個のモデルのうち，説明変数を3個だけ使う286個を$\mathcal{M}$とした結果を表2に示す．$\hat{\alpha}=\langle$afm$\rangle$ だが，相対的なモデルの良さを比較するための p-値 ($P_\alpha^{(L)}$, $P_\alpha^{(M)}$, $P_\alpha^{(B)}$) では，幾つかのモデルで α^* の可能性が示唆されている．これに対し，フルモデルに対する尤度比検定の p-値 ($P_\alpha^{(S)}$) では，$\hat{\alpha}$ を含めすべての $\alpha\in\mathcal{M}$が棄却されてしまう．

13個のうち3個だけを使うという制約は恣意的であり，すべての候補が $P_\alpha^{(S)}$ で棄却されるのは驚くことではない．しかし何らかの事情によりこのような制約下で良いモデルを探す必要がある場合には，AICや，$P_\alpha^{(L)}$, $P_\alpha^{(M)}$, $P_\alpha^{(B)}$ などの指標が役立つ．

4.1 ブートストラップ選択確率

データ $x^{(n)}$ からリサンプリングによって $X^{*(n)}=(X_1^*,\ldots, X_n^*)$ を多数発生させ，各モデルが選択される頻度を推定したものはブートストラップ選択確率と呼ばれ，Felsenstein (1985) によって分子系統樹のトポロジ推定に使われた．これは，真の分布からデータ $x^{(n)}$ を発生させた時に各モデルが選ばれる確率の推定であり，$P_\alpha^{(B)}=$ (モデル α が選ばれた回数)/(リサンプリング回数) である．数値例では，モデル選択にはAICを用いており，リサンプリング回数は 10^4 である．例えば表1の$\langle$abd$\rangle$は，そのうち約1800回程AIC最小になっている．定義より，$\sum_{\alpha\in\mathcal{M}} P_\alpha^{(B)}=1$ である．

本稿ではデータをリサンプリングする代わりに，$\log p_\alpha(x_t|\hat{\theta}_\alpha)$, $t=1,\ldots, n$, $\alpha\in\mathcal{M}$から，対数尤度を直接リサンプリングしている．ただし，Kishino et al. (1990) の正規近似を利用しており，同様の方法は4.3節でも用いた．

しばしば，ブートストラップ選択確率の大きいモデルを集めてモデルの「信頼集合」が作られるが，(2.3) の意味で一番良いモデルとの関係は明確には与えられていない．また，予測分布の似たモデルがいくつかあると，それらで選択の頻度を分けあってしまい，悪いモデルでも似通ったものの少ない方が選択頻度は大きくなるなどの点に注意する必要がある．

4.2 モデル選択検定

Linhart (1988) はAICのサンプリングエラーを評価するために，二つのモデルの最大対数尤度 $L_\alpha^{(n)}(\hat{\theta}_\alpha)$ の差の分散を

$$\hat{\sigma}_{1,2}^2=\sum_{t=1}^{n}(\log p_1(x_t|\hat{\theta}_1)-\log p_2(x_t|\hat{\theta}_2))^2-\frac{1}{n}(L_1(\hat{\theta}_1)-L_2(\hat{\theta}_2))^2 \tag{4.1}$$

で推定し，適当な条件で漸近正規分布に従う統計量

$$T_{1,2}=\frac{\mathrm{AIC}_1-\mathrm{AIC}_2}{2\hat{\sigma}_{1,2}} \tag{4.2}$$

を使ってAICの差の有意性を検定した．ただし，AICは(2.4)を$2n$倍したものとする．$\log p_\alpha(x_t|\hat{\theta}_\alpha)\approx\log p_\alpha(x_t|\theta_\alpha^*)$ と近似すると $L_\alpha^{(n)}(\hat{\theta}_\alpha)\approx L_\alpha^{(n)}(\theta_\alpha^*)$ が n 個の独立な確率変数の和になる事から (4.1) が導かれる．標準正規分布 $N(0,1)$ の分布関数を $\Phi(x)$ とすると，$T_{1,2}>\Phi^{-1}(1-P^*)$ なら近似的に有意水準 P^* で $\mathrm{AIC}_1-\mathrm{AIC}_2$ の期待値は正とみなされる．従ってモデル2の方がモデル1より良いと判断できる．一方，$\mathrm{AIC}_1-\mathrm{AIC}_2>0$ でも $T_{1,2}<\Phi^{-1}(1-P^*)$ なら，モデル2とモデル1の良さに有意な差は無いと判断される．

このLinhart検定は，同じような状況で使われる古典的なCox (1962) の検定とは考え方が全く異なる．モデル1をモデル2に対して検定する場合，Cox検定の帰無仮説は $q(x)\equiv p_1(x|\theta_1^*)$

であるが，Linhart検定の帰無仮説はモデル1の(2.3)(もしくはAICの期待値)が，モデル2のそれ以下になることである．

モデル候補 $\alpha \in \mathcal{M}$ が3個以上の場合には，AIC最小モデル $\hat{\alpha}$ を対照(control)としたLinhart検定が考えられる．すなわち，$\mathrm{AIC}_{\hat{\alpha}} = \min_{\alpha \in \mathcal{M}} \mathrm{AIC}_{\alpha}$ よりも各AICが有意に大きいかを検定する．$T_{\alpha,\hat{\alpha}} > \Phi^{-1}(1-P^*)$ ならモデル α は $\hat{\alpha}$ より悪いと判断される．従って $T_{\alpha,\hat{\alpha}} \leq \Phi^{-1}(1-P^*)$ となる $\alpha \in \mathcal{M}$ を集めると，これはモデルの信頼集合と考えられる．同じことだが，$P_{\alpha}^{(L)} = 1 - \Phi(T_{\alpha,\hat{\alpha}})$ と定義すると，$\{\alpha \in \mathcal{M} : P_{\alpha}^{(L)} \geq P^*\}$ が信頼集合となる．例えば表1で $P^* = 0.05$ とすると，この信頼集合は上位7個のモデルよりなる．しかし，$\hat{\alpha}$ も確率変数なので，この検定の実際の有意水準は P^* にはならない．この問題を解決するのが，次節で述べる方法である．

実は $T_{1,2}$ が $N(0,1)$ に収束するためには，漸近的に $D_1^* = D_2^*$ であるだけでなく，$p_1(x|\theta_1^*) \neq p_2(x|\theta_2^*)$ が必要である(Vuong (1989))．これは3節の図6右のパターンに対応し，もし二つのモデルがネストしているならこの仮定は成立しない．Shimodaira (1997) は，このような場合でも $\hat{\sigma}_{1,2}^2$ にもうひとつ高次の項 $v_{12}/2$ まで含めると，Linhart検定の誤り確率はほとんどの場合に P^* 以下になることを示した．ただし v_{12} は二つのモデルの角度に関係した量で，$0 \leq v_{12} \leq \dim \theta_1 + \dim \theta_2 - 2\dim(p_1(\cdot) \cap p_2(\cdot))$ である．本来なら正規近似よりも，例えばブートストラップなどを利用してより精密な議論をすべきである．しかしモデル選択検定は，特にネストで無いモデルの比較には実用的である．Kishino and Hasegawa (1989) では，分子系統樹のトポロジ推定にこの方法を用いている．

4.3 多重比較による信頼集合

前節では $T_{\alpha,\hat{\alpha}}$ を統計量とする検定を考えた．しかしその導出では確率変数 $\hat{\alpha}$ を定数として扱っていることが問題になる．もし $T_{\alpha,\hat{\alpha}}$ の代わりに T_{α,α^*} が利用できればこの問題は解決する．$\hat{\alpha}$ ではなく，未知の α^* を対照としてAICの差の有意性を検定するために，多重比較 (Hochberg and Tamhane (1987) など) の手法を利用することをShimodaira (1993, 1998a)，下平 (1993, 1995) では提案した．

AIC_{α} と AIC_{α^*} の期待値が等しい事を帰無仮説とした検定をするには，統計量

$$T_{\alpha} = \max_{\beta \in \mathcal{M} \setminus \{\alpha\}} T_{\alpha,\beta} \tag{4.3}$$

が後述する棄却定数 c_{α} より大きいかを見る．$T_{\alpha} > c_{\alpha}$ となるモデル α は，ある $\beta \in \mathcal{M} \setminus \{\alpha\}$ より有意にAICが大きい．従って $\mathrm{AIC}_{\alpha} - \mathrm{AIC}_{\beta}$ の期待値が正と見なされ，結果として $\mathrm{AIC}_{\alpha} - \mathrm{AIC}_{\alpha^*}$ の期待値も正と見なされる．すべての AIC_{β}，$\beta \in \mathcal{M}$ の期待値が等しいと仮定したとき，AIC_{α} が少なくともどれかひとつの AIC_{β} より有意に大きいと判断されてしまう確率が P^* になるように c_{α} を決める．

実際には漸近正規性を利用した近似に基づくMonte-Carlo法で次のように計算する．$|\mathcal{M}|$ 次元の多変量正規分布に従う $U = (U_1, \ldots, U_{|\mathcal{M}|})'$ で，$E(U_{\alpha}) = 0$, $E((U_{\alpha} - U_{\beta})^2) = \hat{\sigma}_{\alpha,\beta}^2$ であるものを疑似乱数を使って多数生成し，$S_{\alpha} = \max_{\beta \in \mathcal{M} \setminus \{\alpha\}} (U_{\alpha} - U_{\beta})/\hat{\sigma}_{\alpha,\beta}$ の分布関数 $F_{\alpha}(x) = \Pr(S_{\alpha} \leq x)$ を近似する．すると，$c_{\alpha} = F_{\alpha}^{-1}(1-P^*)$ である．

もし $T_{\alpha} \leq c_{\alpha}$ ならモデル α と α^* とのAICの差は有意でないとされ，このような α を集めるとモデルの信頼集合 $\hat{\mathcal{T}} = \{\alpha \in \mathcal{M} : T_{\alpha} \leq c_{\alpha}\}$ になる．$\hat{\alpha}$ は α^* の「点推定」なのに対し，$\hat{\mathcal{T}}$ は一種の「信頼区間」になり，漸近正規近似の下で

$$\Pr(\alpha^* \in \hat{\mathcal{T}}) \geq 1 - P^* \tag{4.4}$$

を満たす．この性質は，4.1 節や 4.2 節の方法では一般には満たされず，多重比較による信頼集合の特徴になっている．なお，有意水準 P^* をひとつ決めて $\widehat{\mathcal{U}}$ を与えるよりも，各モデルの p-値を表示した方がより便利である．すなわち

$$P_\alpha^{(M)} = 1 - F_\alpha(T_\alpha) \tag{4.5}$$

を各 $\alpha \in \mathcal{M}$ に計算しておくと，任意の P^* について $\widehat{\mathcal{U}} = \{\alpha \in \mathcal{M} : P_\alpha^{(M)} \geq P^*\}$ がいえる．例えば表 1 で $P^* = 0.05$ とすると，この信頼集合は上位 8 個のモデルよりなる．

ここで述べた方法以外にも，さまざまな多重比較法の適用が考えられる．例えば，(4.3) の代わりに $T_\alpha = \max_{\beta \in \mathcal{M}\setminus\{\alpha\}} w_{\alpha\cdot\beta}(\mathrm{AIC}_\alpha - \mathrm{AIC}_\beta)$ を用い，c_α の計算では $S_\alpha = \max_{\beta \in \mathcal{M}\setminus\{\alpha\}} w_{\alpha\cdot\beta}(U_\alpha - U_\beta)$ を用いる方法でも，(4.4) は満たされる．ただし $w_{\alpha\cdot\beta}$ は適当な定数であり，例えば $w_{\alpha\cdot\beta} = 1$ としてもよい．どのような統計量が実際の問題で良い性能を示すかを調べたり，$\hat{\sigma}^2_{\alpha\cdot\beta}$ を定数と見なすという制約を緩和するのはこれからの研究課題である．

表 2 をみると $P_\alpha^{(L)}$ や $P_\alpha^{(M)}$ の大きさは，必ずしも AIC の大きさの順にはなっていない．例えば AIC の順で 12 番目の 〈klm〉 は比較的大きな p-値になっている．これは，$\mathrm{AIC}_{\langle klm\rangle} - \mathrm{AIC}_{\tilde{a}}$ に対して $\hat{\sigma}_{\langle klm\rangle,\tilde{a}}$ が比較的大きいことによる．単に AIC や p-値を示すだけでは各モデルの良さの程度は分かっても，このようなモデル相互の関係は分かりづらい．このための方法を次節では扱う．

5. 予測分布の相互関係

応用分野における実際のモデル開発には試行錯誤の側面がある．情報量規準の利用によってモデルの適切さを定量的に測ることが可能になり，さらに選択の信頼性を考慮したモデルの信頼集合は，より良いモデルを誤って見落としてしまう可能性を減少させる．しかしそれだけでは不十分で，得られた比較的良いモデルの解釈を行い，それらの相互の関係を総合的に把握し，さらに良いモデルの可能性を示唆するために「予測分布地図」(predictive density map, PDM) を Shimodaira and Cao (1998) は提案している．これは下平 (1993) のモデル地図を一般化したもので，また石黒 (1994) でもなんらかの意味でモデルの地図が書けないかと議論されている．PDM は確率分布の空間における，各モデルによる予測分布の直接の視覚化であり，特に互いにネストしていないモデルの比較に効果がある．このグラフィカルな方法は，例えば共線性の発見など，モデル選択の診断にもつながる．

各モデル $\alpha \in \mathcal{M}$ の予測分布 $p_\alpha(\cdot|\hat{\theta}_\alpha)$ に関して，成分が次のように定義される n 次元ベクトル $\hat{\boldsymbol{\xi}}_\alpha$ を考える：

$$\hat{\xi}_{\alpha,t} = -\log p_\alpha(x_t|\hat{\theta}_\alpha); \qquad t = 1,\ldots, n. \tag{5.1}$$

これはモデル α における各サンプルのエントロピーもしくは符号長である．成分がすべて 1 の n 次元ベクトルを $\mathbf{1}_n$ と書くと，$L_\alpha(\hat{\theta}_\alpha) = -\mathbf{1}_n'\hat{\boldsymbol{\xi}}_\alpha$ や

$$\hat{\sigma}_{\alpha,\beta} = \|\hat{\boldsymbol{\xi}}_\alpha - \hat{\boldsymbol{\xi}}_\beta\|^2 - (\mathbf{1}_n'\hat{\boldsymbol{\xi}}_\alpha - \mathbf{1}_n'\hat{\boldsymbol{\xi}}_\beta)^2/n \tag{5.2}$$

などが言えるので，表 1 や表 2 にあるすべての統計量は $\hat{\boldsymbol{\xi}}_\alpha$, $\alpha \in \mathcal{M}$ とモデルのネスト関係の情報だけから計算できる．それだけでなく，以下で示すように $\hat{\boldsymbol{\xi}}_\alpha$ はモデルの相互関係を理解するのに役立つ．

二つの密度関数 $p_1(x)$ と $p_2(x)$ の Jeffreys 情報量 $J(p_1; p_2)$ は，

$$\int (p_1(x) - p_2(x))(\log p_1(x) - \log p_2(x))\, dx \tag{5.3}$$

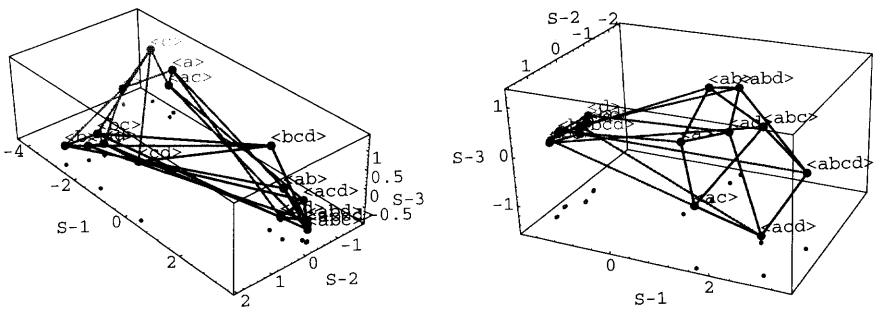

図8. 重回帰モデルの説明変数選択における予測分布地図 (PDM). 4個の説明変数をもつ〈abcd〉と, この部分モデルをあわせて $2^4=16$ 個のモデルの予測分布を, 3個の主成分を使って3次元で表示している. 対数尤度の大きさに応じて点の濃さを変えてある. モデル間をつなぐ線分はネスト関係を表す. [左図] HALD データ. 3個の主成分の累積寄与率 $\rho_3=0.94$. [右図] 新生児の体重データ ($\rho_3=0.94$). 佐和 ((1979), p. 57) より取った. データ数は $n=15$ で, 4個の説明変数がある.

で与えられる. $J(p_1;p_2)=D(p_1;p_2)+D(p_2;p_1)$ の関係があり, これは分布間の隔たりを表している. (5.3) は次のようにも書ける.

$$(1+O(\lambda))\int q(x)(\log p_1(x)-\log p_2(x))^2\,dx \tag{5.4}$$

ただし $\lambda=J(q;p_1)^{1/2}+J(q;p_2)^{1/2}$ は十分小さいとする. この導出は, $p_1(x)-p_2(x)\approx q(x)(\log p_1(x)-\log p_2(x))$ を (5.3) に代入すれば良い. (5.4) を予測分布に適用すると, 二つの予測分布 $p_\alpha(\cdot|\hat\theta_\alpha)$ と $p_\beta(\cdot|\hat\theta_\beta)$ が共に $q(\cdot)$ に十分近いとき,

$$\|\hat{\boldsymbol{\xi}}_\alpha-\hat{\boldsymbol{\xi}}_\beta\|^2\approx nJ(p_\alpha(\hat\theta_\alpha);p_\beta(\hat\theta_\beta)) \tag{5.5}$$

が十分大きな n で言える. 従って, n 次元ユークリッド空間に $\hat{\boldsymbol{\xi}}_\alpha$, $\alpha\in\mathcal{M}$ をプロットしたものは, 確率分布の空間における予測分布の相対的な位置関係を近似的に表している. この図を予測分布地図 (PDM) と呼び, 実際に描画するときは主成分分析 (PCA) を用いて低次元に射影する.

例5の PDM を図8左に示す. 前節の方法で良いモデルと見なされたものは, 右の方に集まっていて, 第1主成分 (S-1) は $\mathbf{1}_n'\hat{\boldsymbol{\xi}}_\alpha$ を反映していることが分かる. 特に, 〈abd〉と〈abc〉はフルモデル〈abcd〉とほとんど見分けがつかず, これらは予測分布の意味で, ほぼ同じ結果をもたらす. 逆に言うと, 与えられたデータからそれらの違いを示すことが難しい. 実は HALD データでは4個の説明変数の和はほとんど一定であり, PDM が縮退しているのはこの共線性の結果である. 一方, 図8右は別のデータの PDM で HALD データのような縮退は見られない. 比較的良いと考えられる z_1 を含む $2^3=8$ 個のモデルの並び方より, z_1 が与えられたときの残りの3個の説明変数の説明力が同じくらいで相互にあまり相関がないことが読み取れる.

正規回帰の場合は, n 次元ベクトル $\hat{\boldsymbol{y}}$ の成分を $\hat y_t=\sum_{i=1}^m\hat\beta_i z_{i,t}$ とすると, $\hat{\boldsymbol{y}}_\alpha/\hat\sigma_m$ は各モデルの説明変数の張る線形部分空間への線形射影もしくは最小2乗法の解として, 幾何的に理解できることが良く知られている. そして, $\hat{\boldsymbol{\xi}}_\alpha$ と類似の PDM を生成することが示せる. (5.1) の利点は, 正規回帰以外の一般の確率モデルに, 同様の幾何的な解釈を直接可能にする点にある. PDM における $\|\hat{\boldsymbol{\xi}}_\alpha-\hat{\boldsymbol{\xi}}_\beta\|^2$ は近似的に, 次のような解釈ができる.

1. (5.5) で与えたように, 予測分布間の隔たり : $nJ(p_\alpha(\hat\theta_\alpha);p_\beta(\hat\theta_\beta))$.
2. 十分大きな n では (5.2) の第2項は無視できるので, 対数尤度の差の分散 : $\hat\sigma^2_{\alpha,\beta}$.

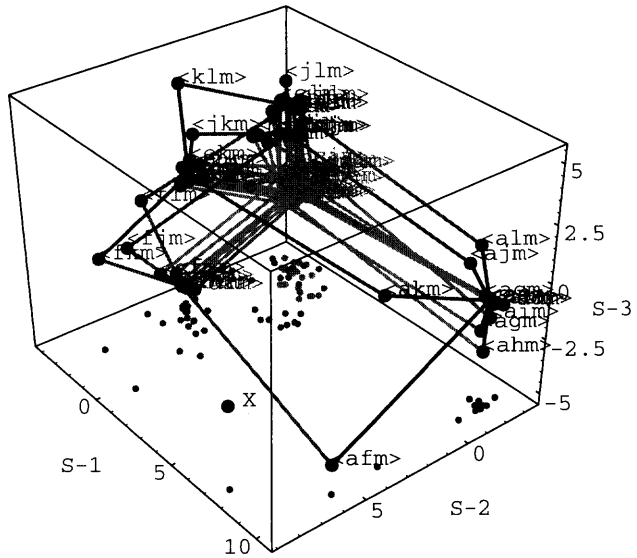

図 9. BOSTON データの PDM ($\rho_3 = 0.72$). 13 個の説明変数のうち 3 個を使うモデルで，特に z_{13} を含む 66 個のモデルに加え，それらの部分モデルが表示されている．また，13 個すべて使うフルモデルも X で示されている．

3．特にネスト $p_\alpha(\cdot) \subset p_\beta(\cdot)$ の関係があるときは，対数尤度のテーラ展開より，対数尤度比統計量：$2 \times \mathbf{1}_n'(\hat{\boldsymbol{\xi}}_\alpha - \hat{\boldsymbol{\xi}}_\beta)$.

このような解釈が正当化される状況として，Simodaira (1997) では，n と共に真の分布を各モデルに $O(n^{-1/2})$ のオーダで近づけ，一般の確率モデルの MLE を正規回帰に還元する方法を議論している．

図 9 は例 6 の PDM である．〈afm〉と〈akm〉がフルモデル X に比較的近く，これらが表 2 で 1 番と 2 番のモデルになっている．しかしそれでも，図 9 の目盛りを見ると分かるように X からの距離は遠く，(解釈-3 より) 対数尤度比統計量が大きくなって $P_\alpha^{(S)}$ では棄却される．表 2 で 3 番から 11 番までのモデルは図 9 の右の方に集まっていて，(解釈-1 より) これらの予測分布は互いに似ている．これに対し，12 番から 16 番のモデルも一群をなしていて，11 番以前の幾つかのモデルよりはかえって p-値が大きい．これは，〈afm〉など上位のものから遠くに離れているから (解釈-2 より) AIC の差が有意でなくなるためである．3 番から 11 番までのものと，12 番から 16 番までのものは，全体として同程度の予測の良さを示しているが，これら 2 群は互いに離れているので，結果として得られる予測のふるまいは大きく異なり，どちらかを選択す

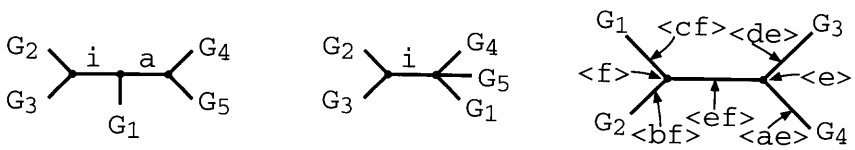

図 10. ここでは G_5 の付く枝に系統樹の根があると考える．[左図] トポロジ $((G_1, (G_2, G_3)), G_4)$ は，〈ai〉と書く．[中図] その枝 a を縮めて 0 にすると部分モデル $(G_1, (G_2, G_3), G_4)$ が得られ，これを〈i〉と書く．さらに枝 i を縮めて 0 にすると，星形トポロジ〈〉が得られる．[右図] $\{G_1, G_2|G_3, G_4\}$ を含むいくつかのトポロジ．G_5 が付く場所を矢印で示した．

る必要がある場合は注意が必要である．この様な幾何的な解釈は，以下に述べる分子進化のモデルでも同様に行える．

例7．（分子進化系統樹の推定） 22種の脊椎動物のミトコンドリアDNAを分析して得たアミノ酸シーケンスデータを用い，Cao et al. (1998) は四足動物の起源を議論している（長谷川，岸野 (1996)）．サイト数（＝シーケンスの長さ）は $n = 3274$ である．ここでは，5個のグループ：G_1＝四足動物（15種），G_2＝肺魚，G_3＝シーラカンス，G_4＝硬骨魚類（4種），G_5＝ヤツメウナギ，の間の系統関係（＝トポロジ）に興味がある．この5個のグループの作る（無根）系統樹のトポロジは15個あり，これらのトポロジは，次の10個の“split”（＝分割）の組合わせで表現できる：a $= \{G_1, G_2, G_3|G_4, G_5\}$，b $= \{G_1, G_3, G_4|G_2, G_5\}$，c $= \{G_2, G_3, G_4|G_1, G_5\}$，d $= \{G_1, G_2, G_4|G_3, G_5\}$，e $= \{G_1, G_2|G_3, G_4, G_5\}$，f $= \{G_3, G_4|G_1, G_2, G_5\}$，g $= \{G_2, G_4|G_1, G_3, G_5\}$，h $=$

表3． 脊椎動物データのAICと p-値（進化速度・不変モデル）．

順位	α	AIC_α	$P_\alpha^{(L)}$	$P_\alpha^{(M)}$	$P_\alpha^{(B)}$	$P_\alpha^{(S)}$
1	<bf>	113877	–	1.000	0.953	0.000
2	<cf>	+80.82	0.023	0.120	0.021	0.000
3	<bj>	+85.98	0.022	0.114	0.020	0.000
4	<bh>	+124.82	0.000	0.003	0.000	0.000
5	<ef>	+139.41	0.000	0.000	0.000	0.000
6	<ci>	+141.73	0.010	0.062	0.006	0.000
7	<cg>	+195.38	0.000	0.002	0.000	0.000
8	<de>	+223.18	0.000	0.000	0.000	0.000
9	<ij>	+231.38	0.000	0.001	0.000	0.000
10	<dj>	+232.40	0.000	0.001	0.000	0.000
11	<ai>	+242.50	0.000	0.000	0.000	0.000
12	<ae>	+253.26	0.000	0.000	0.000	0.000
13	<dg>	+267.05	0.000	0.000	0.000	0.000
14	<ah>	+303.83	0.000	0.000	0.000	0.000
15	<gh>	+313.46	0.000	0.000	0.000	0.000

注）MLEはプログラムパッケージPAML (Yang (1997)) で計算した．$P_\alpha^{(S)}$ はPDMの10次元空間表現より得られた．

表4． 脊椎動物データのAICと p-値（進化速度・変化モデル）．

順位	α	AIC_α	$P_\alpha^{(L)}$	$P_\alpha^{(M)}$	$P_\alpha^{(B)}$	$P_\alpha^{(S)}$
1	<cf>	104790	–	0.946	0.516	0.000
2	<bf>	+5.46	0.385	0.794	0.354	0.000
3	<ci>	+23.83	0.135	0.443	0.108	0.000
4	<ef>	+26.31	0.036	0.178	0.002	0.000
5	<cg>	+44.41	0.007	0.041	0.000	0.000
6	<bj>	+45.73	0.074	0.226	0.015	0.000
7	<de>	+55.19	0.015	0.103	0.002	0.000
8	<bh>	+60.76	0.018	0.032	0.000	0.000
9	<ij>	+62.76	0.022	0.145	0.001	0.000
10	<dj>	+64.17	0.020	0.132	0.001	0.000
11	<ae>	+64.40	0.004	0.036	0.000	0.000
12	<ai>	+68.56	0.012	0.078	0.000	0.000
13	<dg>	+78.18	0.003	0.027	0.000	0.000
14	<ah>	+88.12	0.001	0.008	0.000	0.000
15	<gh>	+90.82	0.000	0.004	0.000	0.000

注）進化速度のサイト間の不均一性はガンマ分布の離散近似 (Yang (1996)) でモデル化し，PAMLでMLEを求めた．

$\{G_1, G_3|G_2, G_4, G_5\}$，i $=\{G_2, G_3|G_1, G_4, G_5\}$，j $=\{G_1, G_4|G_2, G_3, G_5\}$．それぞれの split には系統樹の枝を対応させて考え，その枝の長さはパラメタである．例えば，〈ai〉は a と i の組合わせで表現される（図 10）．15 個のトポロジはすべてこの様に書けるので，系統樹トポロジの選択は例 1 と本質的に同じであり，枝の長さが回帰係数に相当する．ただし，二つの split の組合わせがサイクルのないグラフ（系統樹）になるためには次のような制約を満たす必要がある：二つの split を $\{x|x^c\}$ と $\{y|y^c\}$ の様に書くと，$x \cap y$，$x^c \cap y$，$x \cap y^c$，$x^c \cap y^c$ のひとつが空集合になる必要がある．

系統樹トポロジをひとつ定め，それにそってシーケンスデータがマルコフモデルで変化すると仮定することにより，確率モデル $p_\alpha(\cdot)$ が得られる．ここでは 15 個のトポロジを考え，これらに対応する 15 個のモデルの良さを評価した結果を，表 3 に示す．$P_\alpha^{(M)}$ によれば，〈bf〉，〈cf〉，〈bj〉，〈ci〉などが良いモデルと見なされる．他の 11 個のモデルが α^* である可能性は極めて小さいと示唆される．

次にマルコフモデルを改良して，サイト間の進化速度の不均一性を考慮して解析した結果を表 4 に示す．増えたパラメタの自由度は 1 だが，AIC の値は 9×10^3 も減少していて，モデルの改良は効果があったと言える．ところが，トポロジ間の AIC の差は減少し，信頼集合はかえって大きくなってしまった．これはどういうことだろうか？

図 11 に 2 種類のマルコフモデルの PDM を示す．全体としてこれらは似通っていて，マルコフモデルの改良では，トポロジ相互の関係はあまり変わらなかったことを示している．ただし，各軸のスケールが約半分に減少していて，トポロジ間の分離が悪くなってしまった．これは，サイト間の進化速度の変化を許したため，実質的な有効サイト数が減ったためと考えられる．

PDM 中の X は，10 個の split をすべて使ったフルモデルを表している．これは系統樹トポロジには対応せず，あえて言うならネットワークトポロジに対応している．このフルモデルの MLE を直接求めることは一般に困難であり，ここでは PDM を正規回帰モデルと近似的に見なして，図中の 15＋11＝26 個のトポロジの MLE から線形代数の方法で求めた．$P_\alpha^{(S)}$ はこの X を使って計算していて，すべてのトポロジがフルモデルに対して棄却されることを示唆している．図 11 のすべてのトポロジは X から遠くに離れていることからも，このことが読み取れる．この結果は交配や組み替えの可能性を示唆していると解釈もできるが，ミトコンドリアデータの場合は生物学的に支持されないだろう．

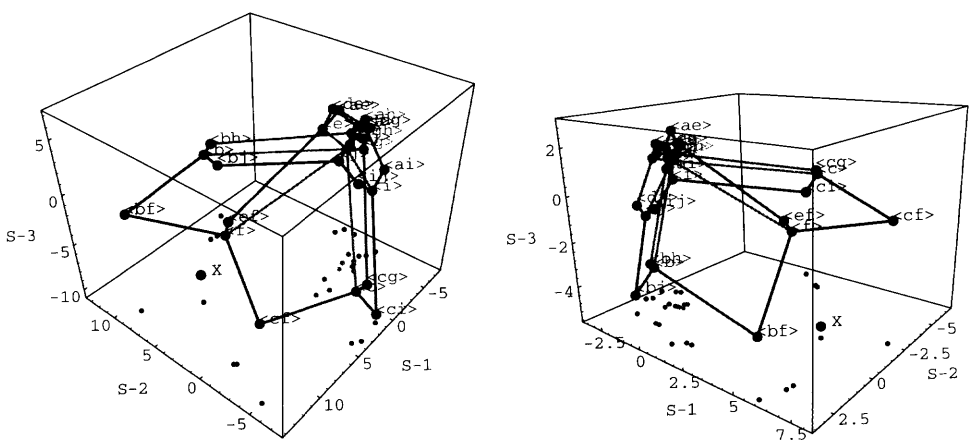

図 11．脊椎動物データの PDM．［左図］進化速度・不変モデル（$\rho_3 = 0.72$），［右図］進化速度・変化モデル（$\rho_3 = 0.77$）．15 個の系統樹トポロジの他に 11 個の部分モデルと 1 個のフルモデルも表示した．

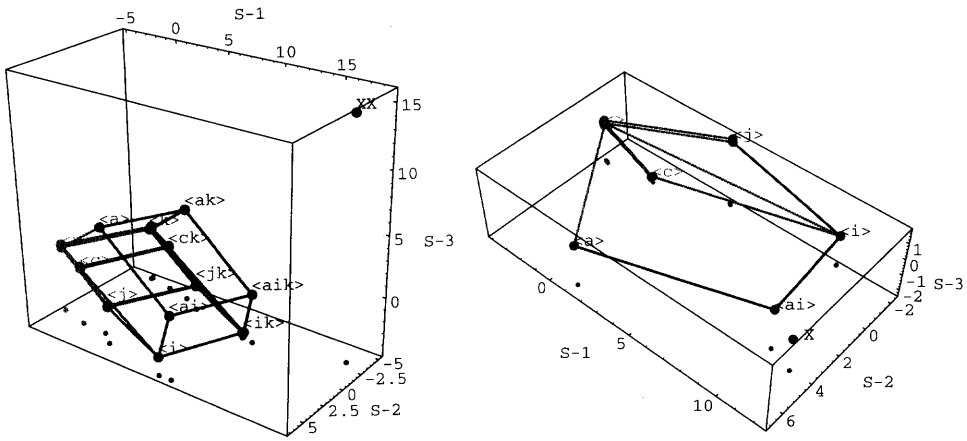

図12. 類人猿データのPDM. [左図] 多項分布のパラメタ推定点XXも表示 ($\rho_3 = 0.97$), [右図] フルモデルXを表示 ($\rho_3 = 0.98$). 枝の長さは非負という制約の結果, ほとんどのトポロジはその部分モデルに縮退している. 左図の因子kは, マルコフモデルの改良に伴うパラメタを表す. 右図ではHKYモデル (k無し, Hasegawa et al. (1985)) を使ったが, TNモデル (k有り, Tamura and Nei (1993)) を使ってもほとんど同じPDMが得られた.

対照的な例として, ヒトの起源を調べるために使われた5種の類人猿のDNAデータ (Horai et al. (1995) ; Adachi and Hasegawa (1996)) を解析したPDMを図12に示す. ただし, G_1=ヒト, G_2=チンパンジー, G_3=ボノボ, G_4=ゴリラ, G_5=オランウータンである. 表は省略するが, AICとp-値のいずれでも, トポロジ〈ai〉だけが明確に支持された. また, $P^{(S)}_{\langle ai\rangle} = 0.039$ であり, PDMでも〈ai〉がXに近いことが読み取れる. ここで注意するのは, 図12左で示されているように, 制約の全くない多項分布のモデルに対しては, どのトポロジも棄却されることである. 従って, 仮定したマルコフモデルはmisspecificationが大きく, 当てはまりの良さの規準では棄却される.

候補となるトポロジの張る空間にXXを射影したものがXであると見なせるので, マルコフモデルをトポロジ間の比較に関する成分とそれに直交する成分に分解して考えたとき, 図12右はその前者の当てはまりは類人猿データでは十分良かったことを示す. 一方, 図11の脊椎動物の例ではトポロジ間の比較に関する成分のmisspecificationさえも大きく, さらなるモデルの改良が必要であることを示唆している.

6. おわりに

「確率モデル」を通して現象を解析する「高度情報処理」は, 今後さらなる発展が期待される. 現在様々な分野でデータが多量に蓄積されており, そこには相互に絡み合った複雑な関係が内在する. このような複雑なデータから有益な情報を取り出し, 解釈し, 判断するには, 統計科学で開発されてきた確率モデルによる方法論が, その柔軟性, 拡張性, 一般性において有効である.

さらに, 近年の急速な計算機技術の進歩に支えられて, データ解析に用いる確率モデルの不自然な制約は取り払われつつある. MCMCなどのシミュレーション技法を用いることにより, 便宜的な制約を気にせずに自由にモデルを構築することが許されてきている.

このような背景で, 可能な無数の確率モデルから有効なものを探し出す方法論の重要性はますます高まっている. あらゆる現象に普遍的にあらわれる分布の構造を調べるとともに, 応用

における個別の問題ごとの専門知識を活用したモデル開発が重要である．本稿で議論してきたのは，このような実際のデータ解析におけるモデル開発を支援するための方法論である．

本稿をまとめるにあたっては，九州大学の小西貞則さん，統計数理研究所の伊庭幸人さん，石黒真木夫さん，栗木哲さん，内田雅之さん，長谷川政美さん，曹纓さん，土谷隆さん，江口真透さん，東京大学の川鍋一晃さんとの意見交換が大いに役立った．ここに記して謝意を表したい．

本研究の一部は文部省科学研究費補助金奨励研究 (A) による．

参 考 文 献

Adachi, J. and Hasegawa, M. (1996). Tempo and mode of synonymous substitutions in mitochondrial DNA of primates, *Molecular Biology and Evolution*, **13**, 200-208.

Akaike, H. (1969). Fitting autoregressive models for prediction, *Ann. Inst. Statist. Math.*, **21**, 243-247.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Control*, **19**, 716-723.

赤池弘次，北川源四郎 編 (1994)．『時系列解析の実際 I，II』，朝倉書店，東京．

Amari, S. (1985). *Differential-geometrical Methods in Statistics*, Lecture Notes in Statistics, Vol. 28, Springer, Berlin.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics*, Wiley, New York.

Cao, Y., Waddell, P. J., Okada, N. and Hasegawa, M. (1998). The complete mitochondrial DNA sequence of the shark *Mustelus manazo*: Resolving vertebrate phylogeny with mitochondrial genome sequences when all known methods fail completely, *Molecular Biology and Evolution*, **15**, 1637-1646.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses, *J. Roy. Statist. Soc. Ser. B*, **24**, 406-424.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, 2nd ed., Wiley, New York.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap, *Evolution*, **39**, 783-791.

Findley, D. F. (1991). Counterexamples to parsimony and BIC, *Ann. Inst. Statist. Math.*, **43**, 505-514.

韓太舜 (1990)．情報理論の最近の展開：エントロピーからコンプレクシティへ，システム/制御/情報，**34**, 71-80.

Hannan, E. J. (1980). The estimation of the order of an ARMA process, *Ann. Statist.*, **8**, 1071-1081.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression, *J. Roy. Statist. Soc. Ser. B*, **41**, 190-195.

長谷川政美，岸野洋久 (1996)．『分子系統学』，岩波書店，東京．

Hasegawa, M., Kishino, H. and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *Journal of Molecular Evolution*, **22**, 160-174.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*, Wiley, New York.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. and Takahata, N. (1995). The recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs, *Proc. Nat. Acad. Sci. U.S.A.*, **92**, 532-536.

石黒真木夫 (1994)．AIC はなぜ役にたつのか？，応用数理，**4**, 125-138.

Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC, *Ann. Inst. Statist. Math.*, **49**, 411-434.

Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea, *Journal of Molecular Evolution*, **29**, 170-179.

Kishino, H., Miyata, T. and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts, *Journal of Molecular Evolution*, **30**, 151-160.

北川源四郎，樋口知之 (1998)．予測とモデル，数理科学，**423**, 11-18.

Komaki, F. (1996). On asymptotic properties of predictive distributions, *Biometrika*, **83**, 299-313.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection, *Biometrika*, **83**, 875-890.
久保木久孝 (1993). ベイズ理論・情報理論とモデル選択, 信学技報 (情報理論), **IT92-130**, 15-22.
Linhart, H. (1988). A test whether two AIC's differ significantly, *South African Statist. J.*, **22**, 153-161.
長岡浩司, 菊池 靖, 池田浩二 (1992). モデル選択における一致性と予測誤差について：AIC と MDL の比較, 第15回情報理論とその応用シンポジウム予稿集, 369-372, 情報理論とその応用学会.
Rissanen, J. (1987). Stochastic complexity, *J. Roy. Statist. Soc. Ser. B*, **49**, 223-239.
坂元慶行, 石黒真木夫, 北川源四郎 (1983).『情報量統計学』, 共立出版, 東京.
佐和隆光 (1979).『回帰分析』, 朝倉書店, 東京.
Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Statist.*, **6**, 461-464.
Shibata, R. (1981). An optimal selection of regression variables, *Biometrika*, **68**, 45-54.
柴田里程 (1988). 変数選択理論の現状, 数学, **36**, 344-352.
Shibata, R. (1989). Statistical aspects of model selection, *From Data to Model* (ed. J. C. Willems), 215-240, Springer, Berlin.
Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection, *Statist. Sinica*, **7**, 375-394.
下平英寿 (1992a). 確率モデルに基づく認識と学習 ── 双対座標を用いた幾何学的アプローチ, 修士論文, 東京大学 計数工学科.
下平英寿 (1992b). 不完全観測対数線形モデルにおける新しい情報量規準, 第14回シンポジウム講演予稿集, 41-46, 応用統計学会.
Shimodaira, H. (1993). A model selection procedure based on the information criterion with its variance, Tech. Report, METR 93-16, University of Tokyo, Japan.
下平英寿 (1993). モデルの信頼集合と地図によるモデル探索, 統計数理, **41**, 131-147.
Shimodaira, H. (1994). A new criterion for selecting models from partially observed data, *Selecting Models from Data : AI and Statistics IV* (eds. P. Cheeseman and R. W. Oldford), Chapter 3, 21-30, Springer, Berlin.
下平英寿 (1995). 統計的モデル選択の研究 ── モデルの信頼集合の構成 ──, 博士論文, 東京大学 計数工学科.
Shimodaira, H. (1997). Assessing the error probability of the model selection test, *Ann. Inst. Statist. Math.*, **49**, 395-410.
下平英寿 (1997). ベイズ的方法における MCMC の利用, 第19回シンポジウム講演予稿集, 1-10, 応用統計学会.
Shimodaira, H. (1998a). An application of multiple comparison techniques to model selection, *Ann. Inst. Statist. Math.*, **50**, 1-13.
Shimodaira, H. (1998b). Improving predictive inference under covariate shift by weighting the log-likelihood function, Research Memo., No. 712, The Institute of Statistical Mathematics, Tokyo.
Shimodaira, H. and Cao, Y. (1998). A graphical technique for model selection diagnosis, Research Memo., No. 680, The Institute of Statistical Mathematics, Tokyo.
Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *J. Roy. Statist. Soc. Ser. B*, **39**, 44-47.
竹内 啓 (1976). 情報統計量の分布とモデルの適切さの規準, 数理科学, **153**, 12-18.
竹内 啓 (1983). AIC 基準による統計的モデル選択をめぐって, 計測と制御, **22**, 445-453.
竹内 啓 編 (1989).『統計学辞典』, 東洋経済新報社, 東京.
Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees, *Molecular Biology and Evolution*, **10**, 512-526.
Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica*, **57**, 307-333 (STMA V31 0456).
White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica*, **50**, 1-26.
山西健司 (1996). 確率的コンプレキシティと学習理論, オペレーションズ・リサーチ, **41**, 379-386.
Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses, *Trends in Ecology and Evolution*, **11**, 367-372.
Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood, *CABIOS*, **13**, 555-556.

Recent Developments in Model Selection Theory

Hidetoshi Shimodaira

(The Institute of Statistical Mathematics)

Data analysis based on stochastic model has been shown useful in many application fields. However, it is often difficult to specify a unique good model from prior knowledge, and so we need a methodology for selecting models from data. Akaike gave the information criterion to evaluate the model in terms of prediction, and he advocated the importance of modeling in data analysis. Up to now, several kinds of information criteria have been proposed in literature, and we have to choose an appropriate one according to our purposes and the situations. In this article, we discuss the derivations of information criteria for several inference schemes. We also make some comments on the consistency of model selection. The issue of consistency concerns the limit of large sample size, but the sample size is finite in actual applications. Thus, it is important to consider the sampling error of the information criterion to evaluate the reliability (or uncertainty) of model selection. Methods such as the bootstrap selection probability, the model selection test, and the multiple comparisons of models are discussed for assessing the reliability of model selection. Further, we give a graphical method to visualize the relative locations of predictive densities for exploratory model building. Illuminating examples from variable selection in multiple regression as well as practical examples from the evolutionary tree reconstruction are given to illustrate the methodology.

Key words: Information criterion, AIC, predictive density, variable selection, Bayes model, multiple comparisons.